Information Densification of Social Constructs via Behavior Analysis of Social Media Users – A Study on Twitter

by

Diogo F. Pacheco

Master of Computer Engineering
University of Pernambuco
2008

Bachelor of Computer Engineering
University of Pernambuco
2006

A Dissertation
submitted to Florida Institute of Technology
in partial fulfillment of the requirements
for the degree of

Doctorate of Philosophy
in
Computer Sciences

Melbourne, Florida
December, 2017
We the undersigned committee
hereby approve the attached Dissertation

Information Densification of Social Constructs via Behavior Analysis of Social Media Users – A Study on Twitter by Diogo F. Pacheco

Ronaldo Menezes, Ph.D.
Professor
School of Computing
Committee Chair

Fernando Buarque de Lima Neto, Ph.D.
Associate Professor
University of Pernambuco, Brazil
Co-Advisor

Nezamoddin Nezamoddini-Kachouie, Ph.D.
Assistant Professor
Mathematical Sciences
Outside Committee Member

Heather Crawford, Ph.D.
Assistant Professor
School of Computing
Committee Member

Philip Bernhard, Ph.D.
Associate Professor and Program Chair
School of Computing
Committee Member
ABSTRACT

Title: Information Densification of Social Constructs via Behavior Analysis of Social Media Users – A Study on Twitter

Author: Diogo F. Pacheco

Advisor: Ronaldo Menezes, Ph.D.

Co-Advisor: Fernando Buarque de Lima Neto, Ph.D.

We live in a digital era where everyday activities are increasingly being replaced by online interactions. In addition, technology advances and data availability are changing the way we expand our knowledge about ourselves, society, and the environment. The increasing availability of data, especially social media data, has called the attention of researchers, and we have been witnessing an outbreak in studies relying on this rich source of information. However, most social media research is tuned to improve the outcomes of specific problems. Therefore, the reuse of techniques used in different areas is limited to data specialists. We propose a straightforward data-driven methodology to perform exploratory analysis of social media data by processing the unstructured stream of social data into user characterization. Emergent collective behaviors are obtained by aggregating individual characterizations. The structured representations are analyzed using Statistics and Data Science techniques. The results highlight the methodology generalization capacity, since we apply it in three different domains: (i) sports, characterizing football supporters; (ii) culture, characterizing languages; and (iii) health, characterizing organ donation awareness. Finally, the knowledge extracted from these applications (experience) serve as input to further research; we propose a measure for social disorganization using the diversity of supporters in a region, and we show language network centralities as proxy for quality of life.
Table of Contents

Abstract iii

List of Figures ix

List of Tables xv

Acknowledgments xvii

Dedication xix

1 Introduction 1

1.1 Manifesto ........................................ 5

2 Background 9

2.1 Classical Studies .................................... 10

2.2 Computational Social Science .......................... 13

2.3 Social Media Data ................................... 18

2.4 Methods, Biases, and Pitfalls ........................ 21

3 Proposed Methodology 25

3.1 Data Collection ...................................... 26

3.1.1 Twitter ........................................ 27

iv
3.1.2 Validation ......................................................... 28

3.2 Characterizing Through Social Interactions ...................... 29

3.2.1 Characterizing Users ........................................... 30

3.2.2 Aggregated Characterizations ................................ 32

3.2.3 Validation ........................................................ 34

3.3 Analyzing the Characterizations .................................. 38

3.3.1 Ranking .......................................................... 39

3.3.2 Highlighting ..................................................... 40

3.3.3 Visualizing ....................................................... 41

3.3.4 Grouping ........................................................ 43

3.3.5 Connecting ....................................................... 44

3.4 Experience ........................................................... 45

4 Learning from Social Media Data .................................... 46

4.1 Characterizing Football Supporters ................................. 47

4.1.1 Motivation ....................................................... 47

4.1.2 Experiment Settings ............................................ 49

4.1.2.1 Methodology Instantiation ............................... 49

4.1.2.2 Characterizing Supporters and Clubs ................. 49

4.1.2.3 Datasets ...................................................... 51

4.1.3 Results .......................................................... 52

4.1.3.1 Ranking ....................................................... 52

4.1.3.1.1 BSA Ranks .............................................. 53

4.1.3.1.2 EPL Ranks .............................................. 55

4.1.3.2 Clubs as Distributions .................................... 57

4.1.3.3 Clustering Clubs ............................................ 58
5 Learning from Characterizations

5.1 Estimating Social Disorganization

5.1.1 Motivation

5.1.2 Experiment Settings

5.1.2.1 Characterizing Regions

5.1.2.2 Datasets

5.1.2.3 The Entropy of Football Supporters

5.1.3 Results

5.1.4 Discussion

5.2 Indicating Socio-Economic Development

5.2.1 Motivation

5.2.2 Experimenter Settings

5.2.2.1 Language Networks

5.2.2.2 HDI of Languages

5.2.3 Results

5.2.3.1 Languages by HDI

5.2.3.2 Correlation with Network Properties

5.2.4 Discussion

6 Conclusions, Limitations, and Future Works

References

A Validation

A.1 User Validation
A.2 Aggregated Validation .......................... 173

B Supplementary Figures .......................... 180

C Twitter Datasets ............................... 190
# List of Figures

1.1 Overview of the proposed methodology. . . . . . . . . . . . . . . . . . 3

1.2 Word clouds extracted from chapters, summarizing their content. . 6

2.1 Percentage (%) of online adults who say they use the following SNS, from 2012–2015 (extracted from Pew Research report [37]). . . . . 19

3.1 Data collection is the first step of the methodology. . . . . . . . . . 26

3.2 The characterization phase is composed by two steps. First, characterizing users, giving raise to structured data. Then, the aggregation of them generating group-characterizations. . . . . . . . . . . . . . 29

3.3 The analyses phase consist of applying traditional computational methods (e.g., visualization, rank, cluster, and network), using the characterizations as input to extract knowledge from data. . . . . 38

3.4 Visual representation of a normalized contingency matrix $\hat{U}$ sorted by preferred entities. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 40

3.5 Different approaches to represent grouped characterizations visually 42

3.6 The last phase is experience, i.e., to use the prior knowledge, already generated by the process, to delve further knowledge. . . . . . . . 45
4.1 Instantiating the methodology. After choosing the context, define the set of entities and the research questions. Then, plan adequately which group-characterizations and analyses will help you to find the answers. .......................................................... 46

4.2 Distribution of non-zero mentioned scores (\(\hat{u}_{ij}\)) for all users (supporters) characterized in the Brazilian (in blue) and in the English (in red) datasets. .......................................................... 50

4.3 Three plots describing two Twitter datasets: tweets and clubs mentioned per user, and clubs per tweet. ................................. 54

4.4 Analysis for the 2015 “Série A” Brazilian league dataset. On (a), clubs as distributions and on (b) and (c) clustering approaches. . . 59

4.5 Analysis for the 14/15 Premier league dataset. On (a), clubs as distributions and on (b) and (c) clustering approaches. .............. 60

4.6 Minna Sundberg’s illustration of the language tree, focusing on the Indo-European and Uralic branches [175]. ......................... 66

4.7 Comparing the correlations between pairs of datasets at different levels of locality. .......................................................... 73

4.8 Measuring the locality effect of datasets on language relationships. . 74

4.9 Number of users per language in the Olympics dataset calculated using Equation 3.11. .......................................................... 75

4.10 Revealing language relationships via network. .......................... 77

4.11 A sample of language characterizations from Olympics16 dataset. . 78
4.12 The locality effect captured by language distributions. Languages characterized in rows—English, French, and Japanese, respectively; and across city-level datasets. For instance, the second language of English speakers is Japanese, Spanish, or French if you live in Tokyo, NYC, or Paris, respectively. 79

4.13 Revealing language relationships via clusters. The language hierarchical clusters for datasets at different locality-levels: (a) Tokyo, (b) The UK, and (c) Olympics. 80

4.14 The set of keywords used to collect tweets related to organ donation awareness is the Cartesian product of Context and Subject words. 90

4.15 Dataset information shown as histograms in log scale. (a) The number of users per organ is a proxy for the “popularity” of each organ in Twitter conversations. (b) Number of organs mentioned per user (blue) and per tweet (red). Organs are more likely to be mentioned when tweets are aggregated by user than on solo tweets. 92

4.16 Characterization of the six major solid organs based on Twitter conversations. 93

4.17 Characterization of US states based on attention given to organs by their Twitter users. 95

4.18 US States colored according to excessive conversations about specific organs. 97

4.19 Hierarchical clustering of states based on their similarity with regards to the extent of incidence of specific organ-related conversations. 99

4.20 Cluster of users based on their conversations on Twitter using K-Means and their relative size. 101
5.1 The knowledge acquired by analyzing the initial characterizations can be used in further investigations. .......................... 103

5.2 Plots of UK with inner plot from London area of: (a) crime, (b) regions labeled according to highlighted football clubs, and (c) entropy of football supporters. ................................. 110

5.3 The partial correlation analysis $\hat{\rho}_{EC,U}$ between all crimes and supporters entropy, controlling for number of users, increases as the minimum users increases, but the cover percentage of regions decreases. ................................. 112

5.4 The entropy of regions in the UK is positively correlated with crime rate, and this correlation is higher in more disorganized places, regardless of crime type. ................................. 113

5.5 Our proxy to social disorganization better explains crime rate in disorganized and populated places. ................................. 114

5.6 The significant positive correlation between language centralities and the Human Development Index indicate that the language positions in the network can be co-related to the QoL. Size and color of the languages represent the standard deviation of the estimated HDI$_{\ell}$................................. 126

6.1 Word cloud formed by this whole dissertation. ................................. 129

A.1 Number of users per activity-level group. ................................. 161

A.2 Average number of entities mentioned by users. ................................. 162

A.3 Temporal evolution of hypothetical characterizations based on Spearman correlation. ................................. 164
A.4 Temporal evolution of user characterizations based on Spearman correlation. ............................................................... 164
A.5 Temporal evolution of characterization based on Bhattacharyya distance. ................................................................. 165
A.6 Distribution of mentioned score (Equation 3.3) for the top-5 rank positions for different groups of user. ......................... 166
A.7 Distribution of mentioned score for the top-5 most mentioned entities for the group of users with 292-527 tweets of the Brazilian football dataset. ............................................................... 166
A.8 Football BRA – Temporal median rank evolution of the final top-5 preferred entities per individual. .............................. 168
A.9 Football UK – Temporal median rank evolution of the final top-5 preferred entities per individual. .............................. 169
A.10 Languages – Temporal median rank evolution of the final top-5 preferred entities per individual. ............................... 170
A.11 Organs – Temporal median rank evolution of the final top-5 preferred entities per individual. ............................... 171
A.12 Analytical evidence for stable preferred entities. Mentioned score gap $\Delta$ and the fluctuation $\delta$ plotted as function of the rank for different groups of users and datasets. ............................... 172
A.13 High top-5 rank stability even for tiny user samples. .......... 174
A.14 Coverage based on different sample sizes of users. ............ 176
A.15 Temporal analyses of aggregated characterizations. ............. 177
A.16 Temporal analyses of aggregated characterizations using the Brazilian football dataset. ............................................ 178
A.17 Aggregated characterizations are more stable than individual characterization (compare with Figure A.12). .......................... 179

B.1 Engagement on Brazilian football dataset. .......................... 181
B.2 Ranking the BSA supporters. .......................... 182
B.3 Ranking the EPL supporters. .......................... 183
B.4 Number of users per language in different datasets calculated using Equation 3.11. .......................... 184
B.5 Language characterizations from the Olympics16 dataset. .......................... 185
B.6 Language characterizations from the Tokyo dataset. .......................... 186
B.7 Language characterizations from the New York City dataset. .......................... 187
B.8 Language characterizations from the Paris dataset. .......................... 188
B.9 Relative risk (log) of organs for each US state. .......................... 189

C.1 The term “olympics” translated to 107 languages according to Google Translate. These are the tracking terms to build the Olympics16 dataset. .......................... 190
# List of Tables

4.1 Twitter official accounts from clubs used in football supporters characterization. ........................................ 52

4.2 Some statistics about the Brazilian and English football datasets. ................................................................. 53

4.3 BSA – proportions, number of users, and rank positions for clubs based on (left) popularity $e_j$ and on (right) opposition $a_j$, compared against Datafolha (DF14), Ibope (I14), and Paraná (P13 & P16) great supporters and most hated club. ..................................................... 55

4.4 EPL – proportions, number of users, and rank positions for clubs based on (left) popularity $e_j$ and on (right) opposition $a_j$, compared against objective ranks from SportsMail as great supporters and Mirror15 as most hated club. ..................................................... 57

4.5 Descriptive statistics of 12 datasets identified by their locality level. ............................................................... 71

4.6 Top ten languages ranked in the LN based on $in/out$ degree and weighted degree, and eigenvector centralities. ..................................................... 76

4.7 Statistics of the dataset used in the organ characterization. ................................................................. 91

5.1 Statistics from the football Twitter dataset. ..................................................... 108

5.2 Official Twitter accounts and hashtags of the football clubs considered in our analysis. ........................................ 109

5.3 Statistics, correlations and regression summary per type of crime. ................................................................. 115
5.4 Highest and lowest ranked countries and languages according to their HDI; HDI$_c$ extracted from UN \cite{160} and HDI$_ℓ$, calculated using Equation 5.2. . . . . . . . . . . . . . . . . . . . . . . . . . . . 123

5.5 Top languages ranked in the G20 and Olympics datasets according to in-degree and eigenvector centralities measured from the respectively language networks. . . . . . . . . . . . . . . . . . . . . . . . . . . . 124

5.6 The correlation values between language centralities and HDI$_ℓ$ in our datasets. Only Eigenvector and In-degree centralities have positive correlation with HDI$_ℓ$ at 99\% confidence level. . . . . . . . . . . . . . 125

A.1 Twitter datasets used in the validation process. . . . . . . . . . . 160

A.2 Toy example of the best case mentioning scenario of a user mentioning 4 of 10 entities . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 163
Acknowledgements

This work was partially supported by CAPES Foundation Science Without Borders Brazilian program under grant 0544/13-2 and by the Army Research Office under grant W911NF-17-1-0127-P00001. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

I would like first to thank my committee Dr. Crawford, Dr. Kachouie, and Dr. Bernhard for their time and constructive comments throughout this doctorate process; this dissertation is the result of your kind and careful vigilance.

I do not want to sound cliché, but none of this would be possible without my advisors Dr. Lima-Neto and Dr. Menezes. Fernando believed on my capacity much earlier than I. He risked his collaboration with Ronaldo by offering this position to me even after I had been absent from academia for 5 years. As co-advisor, Fernando was always pushing me out of the comfort zone, requesting intensive reflections regarding the why’s, what’s, and how’s of the scientific method. Ronaldo fascinated me with his availability and countless interesting ideas to make science. He taught me how to become an independent research. Even using Portuguese in most of our conversations, he improved substantially my writing and speaking English skills, always giving feedbacks and recording my presentations. Rapidly, we started sharing also several hours off-work cycling and drinking beer—a friend
for life. A special thanks to Dr. Ribeiro—an informal advisor, a co-author, and a friend—who was always promptly available to advice with any sort of problem—family, math, English, jobs, to name but a few.

Thank you BioComplex lab (current and former) mates for discussions, proof-readings, collaborations, and stand-up meetings, breakfasts, and lunch conversations. All this together helped me to construct a broader comprehension about our world, from politics to religion and everything in between. A special thanks to Hugo, Marcos, and Diego. Hugo has been supporting me even before my admission until today, no mater whether with miscellaneousness, python, or recommendation letters. Marcos has been sharing and pushing his eccentric rumble ideas since our first day here (Aug 11th 2013); for sure, he is the one I have spent more time with on campus talking about research, freewill, God, happiness in the morning, and football. Diego, in addition to research collaborations, has become member of the family—now I have a younger brother!

Finally, but not least, many thanks to my spouse Mariana and my sons Tiago and Filipe. You were the ones suffering on the daily base with my absence, stress, and changes of mood. I hope I have also succeeded as husband and father.
Dedication

To my wife Mariana, who made this utopian journey become true. Since the first invitation in September 2012, she was the one who pushed me harder to say yes to this endeavor. She not only supported my decision, but she has been contributing to my doctorate in every possible aspect since then. Thank you for all abdication, dedication, and care. I love you!

To my sons Tiago and Filipe, who despite being unaware of their importance to my success, they have been fundamental to my mental health. Their daily high dosages of ingenuous true love kept me motivated to never gave up, and at the same time, transformed my doctoral student period in the happiest moments in my life.

Finally, to my parents Pacheco and Geyne, who have been seeding the importance of a good education, since my childhood, in the formation of rational individuals, engaged citizens, and therefore, towards a peaceful and egalitarian society.
Chapter 1

Introduction

Modern life is increasingly containing online interactions. Simple daily activities like shopping, attending courses, or communicating with friends are shifting to the digital world [171, 45, 154]. In addition, technology advances and data availability are changing the way we expand our knowledge about ourselves and about society. All that produce digital footprints which have been used to track habits and preferences, affording personalized advertisement. Such vast collection of digital footprint have been used to predict/influence our future behavior [81, 51, 57, 155, 156]. In this dissertation we examine how much they can tell about our society.

As product of evolution and experiences (nature and nurture), when in society, humans are more often characterized by the latter, i.e., the cultural traits transcend generations [83, 27]. Hence, we argue that people do behave accordingly to their history of life. In other words, can experiences (cultural, social, economic, etc.) shape up people everyday attitudes in such a way to allow them to be categorized? Moreover, if the answer is yes, do social online interactions embed such characterization?
A common practice in societies is to use surveys to sense population's attitudes and opinions about a multitude of topics, or to engage them into public decisions. Surveys and pools are reliable processes derived from statistics and based on probabilistic models. Inevitably, mainly because of lower costs and agility, these methods are also shifting into online formats. Although an increasing portion of the population have access to the Internet—70% of the world’s youth are online—a fraction of the population is still not represented, for instance, about 19%, 59%, and 82% of the population of developed, developing, and less-developed countries do not have access to the Internet, respectively [150, 137]. This fact has motivated studies showing the resultant coverage errors are modest in most cases, but questions related to subgroups in the population (underrepresented online) can be very biased, such as for 65+ age group [71, 72].

The increasing availability of data, especially social media data, has called the attention of researchers, and we have been witnessing an outbreak in studies relying on this rich source information. Currently, most prestigious magazines/journals feature social media data to understand health, sociology, economics, marketing, human mobility, security, to name a few [12, 131, 54, 7, 74, 41, 173, 86, 73, 20]. However, most social media research is tuned to improve the outcomes of specific problems. Therefore, the reuse of techniques used in different areas is limited to data specialists.

In this work, we propose a methodology to help the exploration of social media data by researchers in different areas with distinct backgrounds. The methodology is intended to be a data-driven straightforward exploratory analysis, on ordinary subjects (entities) in social network sites (SNS), for: (i) characterizing entities through people’s online behavior, (ii) determining rankings of similar entities, and
(iii) identifying the relationships between them. As a general approach, we do not expect the use of our methodology leads to huge breakthroughs, but introduces new data sources, and provides insights for further research. Yet, we claim the knowledge obtained through these characterizations is substantial enough to be used as input for other research. Figure 1.1 presents an schematic view of this work.

Due to their worldwide popularity and importance to society, we applied the methodology in three different domains – sports, culture, and health. In addition, their distinctiveness can indicate how versatile our methodology is.

First, we use large Twitter datasets related to football in the UK and Brazil to characterize supporters and clubs, to confront their popularity, and to reveal rivalries. Then, we aggregate supporters by regions to correlate levels of rivalry with non-obvious social behavior such as levels of crime, developing a proxy to social disorganization\(^1\). Second, we explored several datasets to better understand how languages spread worldwide. We discovered that multilingual users tend to use languages based on geographical neighborhood, not based on language syntax.

\(^1\)Social Disorganization is an ecological theory linking crime and anti-social behavior to neighborhood ethnography, developed by the Chicago School [135].
similarities. Moreover, we have found languages’ position in a language network can be used as a proxy for quality of life of their speakers. Finally, we use a Twitter dataset related to organ donation in the US to characterize organs through mentions. The organ characterization reveals uneven co-occurrence of them while aggregating mentions at state level shows discrepancies on which organs are being highlighted. And for a practical use, we grouped users to understand their necessities and similarities.

An important disclaimer is that we are not proposing methods to replace traditional probabilistic-based models, but alternatives to capture similar outcomes more quickly, dynamically, and possibly allowing us to understand social behavior in shorter terms than traditional methods do (e.g., decennial census). Moreover, we are not proposing a methodology to replace Social Science ones. We do believe good science is rarely done by quick questions and answers, but through a mature, consistent, and incremental process of questioning and analyzing data.

This dissertation is organized as follows (see Figure 1.2 for a visual summary representation of chapters):

**Chapter 2** – presents background and related work, beginning with some traditional social theories until the modern use of computational social sciences. We present some features regarding social media data, and we point out potential biases and pitfalls when dealing with social data.

**Chapter 3** – describes the proposed methodology, how we can progress from unstructured data to densify information based on social constructs extraction.

**Chapter 4** – instantiates the methodology in three different domains by characterizing football supporters (Section 4.1), multilingual users (Section 4.2),
and organs donation awareness (Section 4.3).

**Chapter 5** – uses the information generated on Chapter 4 (learning from social media) as input for further research. We estimate social disorganization based on supporters diversity (Section 5.1), and we correlate quality of life with language position in language networks (Section 5.2).

**Chapter 6** – summarizes our contributions, discussing the limitations of the methodology, and pointing the direction of future works.

### 1.1 Manifesto

To wrap up, the chapters of this dissertation can be seen as world clouds in Figure 1.2, and the Dissertation Ph.D. Manifesto is put forward:

**Area** – Data Science and Computational Social Science.

**Motivation** – Data Science has been successfully applied to extract knowledge and to provide insights from structured and unstructured data in many different areas by combining a multitude of techniques, such as data analysis, machine learning, and complex networks. However, current works are tuned towards their final applications from earlier stages, e.g., data collection or modeling, making their contribution hard to be adaptable to other domains, especially by end users.

**Scientific Problem** – Knowledge retrieval through user interactions on Twitter.
Research Question – To what degree can online social media data be used to identify behaviors and relationships from social actors in a meaningful way.
to enable knowledge extrapolation?

Hypothesis –

1. Users can be characterized based on their social activity in relation to a set of entities.
2. These characterizations can be used to extract knowledge about entities relationships and the interplay between entities and society.
3. The new knowledge is valid, and can be used to support further social analyses.

Objectives –

1. To propose a methodology capable of capturing social relations and unveiling new knowledge about social actors, based on the characterization of users and their relationships with entities.
2. To apply the proposed methodology in three different domains.
3. To use the knowledge learned from the methodology applications in further social research.

Expected Results –

1. Automatic extract of social information from entities relationships.

Scientific Contribution – A methodology for knowledge retrieval from social media data based on the interaction of the social actors (users) and a pre-defined set of entities.
Papers published by the end of this dissertation –

1. Characterization of Football Supporters from Twitter Conversations – application of the methodology in football [114] (Section 4.1).

2. Sensing Language Relationships from Social Media – application of the methodology in language [115] (Section 4.2).

3. Characterizing Organ Donation Awareness from Social Media – application of the methodology in organ donation [113] (Section 4.3).

4. Using Social Media to Assess Neighborhood Social Disorganization: A Case Study in the United Kingdom – using knowledge generated (results of the previous publication) in further research estimating social disorganization [112] (Section 5.1).

5. Language as a Measure of Welfare – using knowledge generated in further research correlating quality of life and languages’ importance in a network [134] (Section 5.2).
Chapter 2

Background

Do we need theories to understand human behavior? Is data enough? With the increasing number of research using big data and digital footprints, we have been witnessing people becoming polarized into two groups: doubters and lovers [59]. The former are skeptical and constantly pointing out to privacy and ethical issues [101]. The latter are further divided in two more groups, the realists and the romantics [94]. The first group perceives new technology and data as tools to assert or to adjust old theories, and to develop new ones; but they are aware of biases and spurious correlations presented in data [140]. The romantics, on the other hand, believe the amount and precision of data are enough to disregard theories and interpretations of them, i.e., “data speaks for themselves” [4].

We recognize the importance of traditional socio-theories and their validity until otherwise disprove. We also understand the importance of privacy and ethical issues, especially when dealing with human subjects. In this chapter, we overview some classical socio-theories and consent upon how social studies evolved from studies targeting individuals until the usage of big (social) data. We describe some
characteristics of social media data, and discuss potential bias and pitfalls of their usage. The idea is to reveal social media data as a useful tool to help us understand human behavior in different domains, but also evidencing the lack of standards to characterize populations and places behaviors.

2.1 Classical Studies

Sociology is the study of societies and the consequences from the interplay between individuals. It aims to develop knowledge about social order, social disorder, and social change. The term Sociology was first coined by Auguste Comte in 1838, on his book series The Course in Positive Philosophy, but sociological thoughts and concerns can be traced back to ancient philosopher Plato [146].

Despite having lived over a century ago, classical sociologists are important to Sociology and Computational Social Science, as their ideas are still relevant and can motivate new discoveries. There is no specific set of attributes to determine one as a classical sociologist, but names as Bernard Mandeville (1670-1733), Adam Smith (1723-1790), Auguste Comte (1798-1857), Alexis de Tocqueville (1805-1859), Karl Marx (1818-1883), Émile Durkheim (1858-1917), Max Weber (1864-1920), and Norbert Elias (1897-1990) contributed to set and establish Sociology as science [170]. Due to objectivity and space, we do not give detailed description about their contribution. Nonetheless, we do highlight some directly opposite theories such as “individuals driving society” versus “society driving individuals”; they are considered one of the main lines of thought and are fundamental to Sociology. Thus, understanding them would engage us more efficiently into the socio-characterizations we are proposing in this work.
Focus on Individual –

Some theorists argue society as being not more than a collection of individuals. They believe the only way to understand society is through its individuals. Adam Smith, for instance, described an invisible force holding society together as the result of self-interest social interactions. The *invisible hand*, as he named, is the emergent force from *social regulation* and *social imitation* [144]. In other words, egoistic attitudes could lead to mutual benefits (or balanced relations).

Max Weber, considered the father of the *Methodological Individualism*, said society could not be anything else than just the individuals that compose it. He defined Sociology as the study of *social action* [43]. These actions or behavioral decisions are performed by individuals and can be classified into four ideal types: *goal-rational, value-rational, affective*, and *traditional*. Goal-rational social actions are those that can be justified with rational arguments. Once the goal is defined, the individual should behave in such way to achieve it. Value-rational social actions can be perceived similar to goal-rational ones, however, the goal itself cannot be justified in rational terms. For instance, one who decides to live solitary in a religious seclusion; the sacrifice of such life can only be explained by the value of believing in rewards after life (it cannot be explained rationally). The affective category is when one acts by impulse, without thinking. Finally, traditional social actions are those we do because it is how society does (e.g., eating with silverware). In modern societies, goal-rational actions are expanding while the other types have been suppressed. These changes are consequence of the slow process of individual *rationalisation* that would, ultimately, reach the *disenchantment* view of the
world [169]. That is, a world where scientific knowledge is more valuable than beliefs.

**Focus on Society** –

There were sociologists focusing on the structure; they viewed society as a complex system. Karl Marx suggested society as being defined by individual relations; the individual essence is not in the part, but it is revealed by the network within it. Marx said: “It is not the consciousness of men that determines their being, but, on the contrary, their social being that determines their consciousness.” [92][p.11-12]. For Marxists, individuals hold their beliefs regarding their social position as *class consciousness*. Class is not only defined by economical aspects, but also by social-psychological characteristics perceived within a group.

Émile Durkheim had a even more holistic view from society than Marx. He observed a positive correlation between the division of labor and societal progress [40]. He believed traditional societies were bound together by similarities (they shared similar languages, gods, religious beliefs, etc.). In modern societies, in contrast, as the levels of specialization raised from industrial revolution and globalization, society would be bound as a result of differences, i.e., their specializations and interdependence would create *organic solidarity*, benefiting society as a whole. Thus, *collective consciousness* is the new cement of society [40]. He established Sociology as a new science based on the study of *social facts* [39], to understand group formation, group identity, group cohesion, etc. In contrast to Weber’s social actions, social facts are inherent to society, they can neither be reduced to individual facts,
nor be used to understand the individual level; they are *sui generis*. Neverthe-
less, they are coercive to individuals. Such emergent behavior, when in
group, is the central argument of the present dissertation.

**In Between Individuals and Society –**

There are also the *figurational* sociologists. Norbert Elias, in his book *The
Civilizing Process* [44], rather than following dichotomies ideas, he focused
on the long chain of interdependence between individuals and society, the
long term changes and their dynamics over humankind. He held neither the
existence of society without individuals nor individuals completed isolated
in society. He opposed static solutions in order to investigate the interaction
processes. To Elias, in the process of becoming civilized (an ongoing pro-
cess), individuals learned to adapt their behavior according to their *social
context*. For instance, one often behaves completely different at bars than at
work. Note that this adaptation process does not mean to hide ones’ real
personality, but a truly comprehension about social constraint toward the
development of self-constraints. Such context based behavior is central for
this dissertation.

### 2.2 Computational Social Science

Since the beginning of Sociology, its classical theorists were doubtful about the role
of societies’ structure for their own development, especially for modern societies.
Either with Durkheim’s static ideas (*funcionalism*) or with Elias’ dynamic process
investigations (*figuration process*) the necessity of a broader comprehension of the
social network was evident.
As we already mentioned, there were beliefs about the role of structure to social networks, but they were not directly tackled in the earlier studies in the field. In this section, we delve through the evolution of social networks studies using some well cited works. We show some works creating networks, others characterizing populations in different ways, and yet few attempts to characterize places. As socio-theories matured, new technologies have been developed, and new types of data have become available, we observed an increasingly multidisciplinary interest for “characterizing” society.

If we could draw a timeline under these studies, in the leftmost end, we would have sociologists revealing some interesting network aspects about our pre-digital society (yet lacking statistical rigor). In the rightmost end, a digital world full of online social networks – the big data era. In between, several threads developing the field, such as:

- an increasing understanding of social structure and its impact in our daily lives;

- plenty of theories and methods coming from mathematics, physics, and statistics, giving birth to a new science – the Network Science [106];

- a shift in the object of study from real individuals to their representation at online social networks;

- a constant concern about how representative these online proxies are to our real society, as well as how trustful are the results derived from them.

One of the first social network studies to become famous was the small-world experiment of Stanley Milgram [98]. Indeed, he was not the first researcher to
investigate the increasingly connectivity in the world, but Milgram is more remembered than his precursors such as Karinthy, Kochen, or Pool [6]. Milgram carried out an experiment in order to check how apart from each other people are in the world. He selected a person in Massachusetts as the target and, then, several random sources in Nebraska. The sources had to mail a package to the target following simple instructions: one should only mail people who he/she knew at first name basis, i.e., if one did not know the target, one should have sent the package to a friend he/she thought was more likely to know the target. The package contained a list to keep track of names it has passed through. He found people were apart, on average, by 6 connections; it was the origin for the well-known expression “six degrees of separation”. Despite being social psychologist, in his experiment Milgram explicit investigated the structure of the network. However, his work has some critical issues regarding its statistics, for instance, Milgram only considered in his calculations letters that reached their destination. Nevertheless, his claim about a small world was later on statistically confirmed in different research [31, 168, 75, 34]. The small-world phenomena still motivates many fields from epidemics to communications [174, 126, 166].

Another interesting aspect revealed by exploring society and its structure is social contagion. Christakis and Fowler performed a series of studies to understand how social behavior could be propagated. They used a dataset collected over a 30-year period from Framingham Heart Study, throughout periodically questionnaires and body measurements. First, they analyzed to what extent obesity could be infectious through social network [24]. From the network formed by friends, spouses, parents, and relatives, they constructed observable networks and random networks with similar properties in order to apply longitudinal statistical models.
to check levels of spreading. They found one has 45%, 20%, and 10% more chance of becoming obese if he/she is connected to an obese person by 1, 2, or 3 degrees of separation (among all types of relationship), respectively. Geographical distance appeared to have no influence in this contagious, but sex and friendship direction indeed impact. For instance, if one (ego) names a friend (alter) that become obese, his/her chances of becoming obese too increases 57%; if they are mutual friends, one’s risk increases by 171%; finally, a friendship originated from alter to ego does not influence on ego’s risk of becoming obese. In 2008, they did a similar study addressing smoking cessation behavior finding akin results [25]. In the last study over social-psychological behaviors of this series, Christakis and Fowler tackled spreading of happiness [48]. They also found happiness to be a contagious phenomenon spreading up to three degrees of separation. However, in contrast to previous studies [24, 25], they found happiness to be more sensitive to time and geographical distance.

Dodds et al. developed the *hedonometer*, an online non-invasive tool to measure temporal patterns of happiness by word analysis using Twitter as content provider [33]. At this time, it is not the spreading mechanism that is being investigated, but different insights to measure feelings in a population. In theory, this new source of data (online social networks) can provide richer and less expensive datasets to formulate new social theories or re-validate the current ones. However, in addition to traditional concerns about theories generalization, external validity, among others, the use of online social networks as proxy for our modern society demands further investigations to assure at what extend they are alike [88, 119].

Gilbert and Karahalios, on the other hand, explored online social networks and the possibility of sentiment measurement to investigate the correlation between
anxiety and worry to stock market variations [53]. They used social media data from LiveJournal to built the Anxiety Index, and the Standard and Poor’s (S&P) Index\(^1\) as the stock market indicator. They found feelings extracted from economical unrelated data in social media helped to point the fluctuation directions of market. In fact, anxiety and worry negatively correlates with S&P 500 closing prices.

Motivated by Adam Smith’s ideas, where the wealth would be the result of specialization (division of labor), Hidalgo and Hausmann presented the concept of building blocks of economic complexity [66]. They compared the economy complexity of countries to Lego buckets diversity, where a Lego piece is a capability. Products would require a set of capabilities in order to be develop and countries could only develop products for which they posses all the required capabilities. In addition, capabilities could not be compared just quantitatively, for instance, some of them were quite ubiquitous among countries, while others less-common were used to build more sophisticated products. They proposed the method of reflections in order to measure and understand the structure of the bipartite network of countries and their exported products [66]. In other words, their measures were a proxy to understand countries capabilities (complexity) by depicting diversification (countries) and ubiquity (products). Moreover, this complexity correlates with GDP and “errors” in the correlation are predictive of future growth.

Poblete et al. were pioneers in investigating language similarities on Twitter [125]. They identified the top ten most used languages and the top three languages for the most active countries. Later, Mocanu et al. used a larger Twit-

\(^1\)http://www.investopedia.com/terms/s/sp500.asp
As they focused on geographical locations, they only used geo-tagged tweets, and found Twitter as a reasonable proxy for language demographics.

Blumenstock et al. used new sources of data to provide demographic and economic profiling [10]. They used mobile phone data from Rwanda to predict poverty and wealth, combining cell phone data, real census data, and a phone survey. The survey with mobile data were used to create a predictive model of wealth. However, instead of using directed variables, they combined them in order to get a wealth index. Since correlations between the predicted index and survey questions were very strong, they estimate the whole dataset for wealth with the benefit of using mobile location information. This approach generated wealth information that used to be retrieved in a much more aggregated view.

Somehow, all those mentioned works are underlying evidences of what we put forward in this dissertation.

2.3 Social Media Data

Social networks sites (SNS) such as Twitter (since 2006), Facebook (since 2004), Google Plus (since 2011), LinkedIn (since 2002), and Instagram (since 2010) are increasingly present in our daily activities. For instance, it is common to find people who more frequently use SNS than e-mail or cellphone (voice).

Some of SNS provide APIs for developers to build apps or to gather data, making data collection process less cumbersome and less dependent of web scraping. As a result, the scientific community also engaged on SNS as source of data or as an object of study. For instance, the number of pages returned when querying for
these SNS in Google Scholar\textsuperscript{2} are: Twitter–6 million, Facebook–5 million, Google Plus–4 million, LinkedIn–1 million, and Instagram–0.2 million. Despite being less popular than Facebook (see Figure 2.1), Twitter is leading the “research rank” mainly because of its openness and powerful API.

Figure 2.1: Percentage (%) of online adults who say they use the following SNS, from 2012–2015 (extracted from Pew Research report [37]).

Facebook and Twitter are by far the most common social media sites. They have been used in diverse contexts of physical and social phenomena, ranging from disaster management [76] to interventions promoting changes in health behavior [82]. For instance, Facebook has been used in interventions involving sexual health [14], physical activity [19, 47], and food safety [93]. Similarly, despite its short-length messages, Twitter data has been shown to be very rich helping researchers understand and predict human behavior. Twitter has been used to understand how collaboration emerges under catastrophic scenarios such as earthquakes [79], to track hurricanes [76], to measure the spread of happiness in a country [34], to predict stock-market transactions [11], and to carry out interventions related to weight-loss [159] and smoking cessation [120].

\textsuperscript{2}by July 14\textsuperscript{th} 2016.
Although both Facebook and Twitter provide a rich source of information, most of them are unstructured text, requiring some transformation to be understood and characterized in order for us to get useful information.

**Twitter Datasets**

In this dissertation, we focus on Twitter as source of social media data. Twitter is an online micro-blogging social network which 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 contributed for this growth. Twitter has become a general-content platform, and it has been used to improve marketing capabilities [87], to predict political campaigns [158], to evaluate entertainment engagement [177], as a media outlet [80], and as an utility service tool [76]. Its users are allowed to exchange private messages (*direct messages*), or to post public messages up to 140 characters (*tweets*, *retweets*, and *replies*)\(^3\). In addition, friendship does not require reciprocity, i.e., friendship is directional through *followers* (incoming friends) and *following* (outgoing friends).

The *Twitter Streaming API*\(^4\) provides real time data collection via several methods, for instance querying keywords, tracking users, constraining regions (bounding-box), or filtering languages. A tweet contain a multitude of fields, such as: id, text, created date, coordinates, language, images, URLs, number of retweets, and an encapsulated user object. Users also have their own fields,

---

\(^3\)While writing this dissertation, Twitter expanded its character count to 280 [123], however, the studies presented here were all based on 140-character tweets.

\(^4\)https://dev.twitter.com/streaming/
such as: id, name, language, location, number of friends and followers, and number of tweets.

2.4 Methods, Biases, and Pitfalls

The conflicting theories of Durkheim and Weber can be seen as the two main methodological approaches in sociology, namely: Positivism (quantitative, rational) and Anti-Positivism (qualitative, interpretative). Beyond this classification, there are still different schools of thoughts, situationism, for instance, perceives quantitative and qualitative research complementary to each other [109]. Since the former aims to answer precisely questions such as what, where, when, how often, and/or how long, the later focuses explaining hows and whys [163].

In addition to the type of targeted question aimed to answer, quantitative and qualitative methods might significantly differ from how they acquire data and the type of data itself. Quantitative method depends on numbers, statistics, reproducibility to be considered valid. Therefore, quantitative research often use empirical data such as surveys, group experiments, and tracking individuals. In contrast, qualitative works focus on interviews, participant observation, field notes, analysis of documents, among others [91, 104].

The more recent area of Social Network Analysis (SNA) can be conceptually classified as a quantitative approach due to its objectivity and rigors. Also, it can be described as a framework to social pragmatic studies, providing exploratory and confirmatory tools [109]. Nevertheless, it is based on Graph Theory and Network Science. In this section, we present studies tackling the representativeness of social media data and investigating possible side-effects of having a online society.
Once in the digital world, one can easily claim distance is no longer a problem to form or maintain relationships. Then, the expected number of virtual friends would be much greater than in the “real life”\(^5\). Or in network terms, online social network is denser than offline social networks. Thus, these differences could compromise the application of traditional socio-theories using this data, or could lead to the development of new theories, exclusively, for online societies. Some works have been done in order to reduce this uncertainty, especially using Twitter data.

Huberman et al. observed the increasing usage of social media (such as Facebook, Twitter, LinkedIn) as proxy for real social networks by market, academia, and governments [67]. Meanwhile, they warn about the differences between declared networks, i.e., those connecting followers to followees, and real interaction networks. The denser social network, therefore, hides a sparse interaction-friend network, i.e., those connecting who has directed post to each other at least twice. This finding raises the importance of deep reasoning of network models in order to achieve realistic results mimicking society.

Dunbar’s number theory states that the number of social ties (stable interpersonal relationships) is biologically constrained by the size of neocortex. Dunbar found the limit for humans to be between 100-200 [38]. Gonalves et al. created a network through Twitter’s conversations to check whether this constraint holds after digital era [58]. They found out the cognitive limitation still plays a role in our relations. Moreover, two basic concepts could explain it: time constraint and prioritization among ties.

\(^5\)We will not discuss the philosophical questions regarding which life – online or offline – is in fact the real one.
Takhteyev et al. investigated the importance of geographic distance, country boundaries, language, and frequency of flights on ties formation [149]. They analyzed how such attributes can be used as tie predictor, and if distance still plays role in social ties formation, since previous study claimed the Internet to diminish the impact of distance [16]. Using Twitter data and the number of direct flights between pairs of airports, their findings revealed distance impacting on ties formation in society. Moreover, they showed the number of flights between micro regions as the best tie predictor when compared with distance, country, or language.

Stephens and Poorthuis presented another way to understand geography and information flow in Twitter by integrating physical and social distances [148]. They found Twitter social neighborhoods to be replications of offline social patterns. In addition, closer communities are denser and more effective in spreading information.

In addition to their capacity of being proxy of real life social networks, SNS can be segmented by their purpose. For example, some of them are used for friendship, others for professional networking, while others for marketplace, and so on. Since Twitter is being one of the most used online community not just as content provider, but even as paper’s quality indicator predictor (citations) [152], it is natural to question its purpose. Kwak et al. investigated several aspects of Twitter network such as: basic topology properties, ranking users, trending topics behavior, impact of retweets [80]. Among other things, they found more than 85% of trending topics are news headlines or media related. Although this media role intuitively suggests top propagators to be celebrities or showbiz accounts (i.e., those with greater number of followers), this was not observed in retweet analysis. They evidenced the emergence of collective intelligence where ordinary people dictate
what is spread over the network.

Moreover, Wang et al. investigated how easily people in different social networks could be traced [165]. They revealed people tend to use similar names and keep same friends over multiple applications such as Facebook, Twitter, and Foursquare. This approach could overcome some data limitation due to specificity, however, they adverted security and leakage issues due to cross self-identification or even by friends reverse lookup.

Yet, Tufekci highlighted some important questions related to the usage of social media data in social studies, especially pointing to potential problems of lack of representation and validation [157]. Olteanu et al. propose a theoretical framework as a guideline for researchers. They focus on different sources of bias, such as population and behavioral biases, biases at the source of data, or introduced in the collection process or while analyzing it. In addition, they advert about interpretation challenges, and argue about ethical choices and practices [108].
Chapter 3

Proposed Methodology

Everyday we seem to interact more online than offline. The virtual world is becoming the most preferred place to meet friends, to find a date, to work, to study, among others. The estimation is that 3.9 billion people are using the Internet, more than half of world’s population\(^1\). As more people are becoming connected and stay connected for longer periods, we may question ourselves: are our online identities equivalent to our “real” identities? Can online behavior or digital footprints be used to characterize offline traits? Can we use online social media data to better understand social-economic indicators from our society?

In this chapter, we propose a methodology to characterize social actors (users) and extract hidden information in social media data, based on their relationship to a set of entities. The methodology, as shown in Figure 1.1, can be described in four major phases: data collection (Section 3.1), characterization (Section 3.2), analyses (Section 3.3), and experience (Section 3.4). Most methods suggested on

\(^1\)http://www.internetworldstats.com/stats.htm
analyses phase are not new or unique. The literature is vast and it is constantly evolving [2, 3, 105, 106, 147]. However, we propose a baseline framework we believe is simple to stimulate the use of data science interdisciplinary.

3.1 Data Collection

Any data analysis project has to deal with data, so we must start the description of the methodology by talking about data (see Figure 3.1). There are many types and sources of data, as well as many forms to have access to them. No matter whether one purchased the data, or downloaded it from an accredited source, or if one had to “crawl” it, or built it from surveys or through a curation process; data needs to be checked. This does not mean one has to have complete data, or error-proof data, but one has to do basic verification in the raw data. For instance, (i) checking appropriateness of data, (ii) describing data attributes, (iii) confirming data types, and (iv) performing data transformations when necessary. These trivial tasks can avoid rework, and more important, they can reduce possible misinterpretation of the results at the end of the process.

Figure 3.1: Data collection is the first step of the methodology.

The quality of any research finding is bounded by the quality of the input data. Therefore, any bias intrinsic to the data source has to be also considered
while interpreting the results. In the following sections some major data sources will be commented upon.

### 3.1.1 Twitter

Most of Twitter data is public. Twitter provides an easy and powerful API to access data, making it one of the most used social media platform in science. Twitter splits its API in two major components: REST and Streaming APIs. In general terms, we could say the REST API is used to retrieve tweets from the past while the Streaming is used to retrieve future occurrences.

**Streaming APIs**

They provide forms to monitor (future) tweets. There are several ways to retrieve data using these APIs, but the most common ones are:

- **Sampling the hose** – Twitter gives 1% of its live stream. This is the simplest approach, since one does not need to specify anything.

- **Tracking terms** – one has to define a strategy to collect tweets by tracking specific terms. A term can be words, hashtags, and Twitter accounts (profiles). Twitter limits the number of tweets to 1% of all stream, but depending on the tracking strategy one might be able to get all tweets within the tracking criteria.

- **Using bounding boxes** – this approach allows to collect tweets based on their location (origin) rather than based on content. Given a set of coordinates (bounding box), Twitter uses users’ profile location, metadata of places marked in tweets, and the actual coordinates provided by mobile gadgets, to retrieve the tweets within a place.
REST APIs –

They provide forms to get the most recent and relevant content from Twitter. The major drawbacks are the rate limits constraints enforced by 15-minute time windows and the chronological short limitation; usually, no greater than the past 7 days.

3.1.2 Validation

An important aspect to consider, when creating your own datasets from social media, is to validate the theme or the target subject. That is, you need to check whether the tracking strategies were effective in capturing what they intend to. A strategy may target events that occur on specific times of the day or in certain days of the week, or even happening in specific places. To do that, one should perform some exploration data analysis building plots to visualize how data is distributed. For instance, you can check the hour/daily frequency of your data, or the spatial distribution.

During this exploratory analysis, one can detect unusual events, i.e., those producing spikes in the data. There are several methods to do that, for instance, trivial visual inspection, simple moving average, autoregressive integrated moving average (ARIMA), and exponential smoothing [13, 63]. Knowing them is important to make decisions regarding data cleaning, or to help understanding the results, especially when performing time-series analysis. For instance, having gaps in the data collection for more than a few hours may compromise the results of that day, then it would be preferable to remove the whole day from the dataset. Or some extreme events, from nature (hurricanes, tsunamis, etc.) or not (terrorism attack, economic crashes, wars, etc.) may skew the data; depending on the application,
one may remove part of data of split it into different periods. Figure B.1 exemplifies the impacts of the tragic plane crash killing almost all Chapecoense club’s football players on Nov 28th, 2016.

3.2 Characterizing Through Social Interactions

The second phase of the methodology is the characterization (see Figure 3.2). As shown in Chapter 2, social media data has been used, not only, to reproduce some classical sociology studies, but to develop new sociological theories or somehow to expand our knowledge about some underlying mechanisms of our society. These studies are often designed to solve specific problems, i.e., their models are tuned (parameter settings or feature selection) to optimize results. In other words, these research have been designed to exploit social media data in specific contexts. In this dissertation, on the other hand, we propose a straight-forward data driven approach to explore social media data in different contexts.

Figure 3.2: The characterization phase is composed by two steps. First, characterizing users, giving raise to structured data. Then, the aggregation of them generating group-characterizations.

The generality aspect naturally imposes limitations in the set of features one can choose. An inherent component on social media are the social actors, i.e., the people, the users. People have different backgrounds and beliefs making direct
individual comparisons complex and subjective. Alternatively, we can define reference points and use them as a comparison basis. Throughout this work, we call the set of these reference points as the set of entities. These entities will be used to characterize social actors, and to instantiate the methodology in different context.

**Entity** – anything that can be identified in a social media post.

**Mention** – is the identification of an entity in a post. For example, a word, a sentence, a hashtag, a set of coordinates, an account (profile), a metadata, the result of a function.

**Set of Entities** – a finite set of related entities in a specific context. For example:
- a set of people or brands identified by specific words or by social media accounts;
- a set of sports, food, colors, books, or things identified by words, sentences, or hashtags;
- places retrieved from metadata;
- languages, emotions, or sentiments resulted from function evaluation.

Ultimately, we propose to characterize social actors from social interactions based on these entities, and to group social actors to obtain higher order characterizations.

### 3.2.1 Characterizing Users

We commence by proposing to characterize the degree of attention given by users to entities from the frequency of mentions about an entity in posts online.

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

\[
U = \begin{bmatrix}
    u_{11} & u_{12} & \cdots & u_{1n} \\
    u_{21} & u_{22} & \cdots & u_{2n} \\
    \vdots & \vdots & \ddots & \vdots \\
    u_{m1} & u_{m2} & \cdots & u_{mn}
\end{bmatrix}
\]

\[u_{ij} = \sum_{t \in \tau_i} W_j(t) C_t, \quad \text{(3.1)}\]

\[C_t = \sum_{j=1}^{n} W_j(t) \quad \text{(3.2)}\]

where \( W_j(\cdot) \) is a function that returns the number of mentions entity \( j \) receives in a post, and \( C_t \) is the total number of mentions in that post (to all entities).

Finally, to ensure all users are treated equally regardless the number of posts they send, we normalize the rows of the matrix so they sum to 1. This is an important aspect of our characterization – it is centered on the user. Thus, our
approach is less likely to be biased by heavily active users or bots, even though the volume of information generated by these exceptional users is high, they contribute to the collective behavior as any other user. We are not advocating all users have the same influence in the network, but our model does not intend to incorporate prior information, i.e., it is unbiased. 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:

$$\hat{u}_{ij} = \frac{u_{ij}}{\sum_{j=1}^{n} u_{ij}}. \quad (3.3)$$

The mentioned scores are normalized by the number of tweets from each user. Thus, we assume tweets from different users can contribute differently to their attention, but for a particular user all of his/her tweets have the same importance. Therefore, a user is characterized in relation to entities as a distribution of mentioned scores, i.e., rows of the normalized contingency matrix $\hat{U}$. Finally, the sum of all elements in $\hat{U}$ is equal to the total number of users, or formally

$$\sum_{i,j=1}^{m,n} \hat{u}_{ij} = m. \quad (3.4)$$

### 3.2.2 Aggregated Characterizations

Usually societies demand polices, actions, laws, etc. based on groups’ necessities. Since individual characterization is not enough to describe a group (too specific), people are commonly identified by categories or communities they are considered being part of. For instance, one could create stereotypes for clubs based on their supporters, for universities based on their students, for states based on their citi-
zens, or for brands based on their customers.

We define a $g \times n$ group-characterization matrix $K = [k_{bj}]_{g \times n}$, where the $b^{th}$ row is the mean of the attention given by all users in the $j^{th}$ group to all entities. Similarly, to users, a group is characterized in relation to entities as a distribution of mentioned scores, but indeed as a collective behavior of its within users. Groups, i.e., rows of matrix $K$, are given by:

$$k_j = \frac{\sum_{i=1}^{m} p_{ij} \hat{u}_i}{\sum_{i=1}^{m} p_{ij}}.$$

(3.5)

$$P = \begin{bmatrix}
p_{11} & p_{12} & \cdots & p_{1g} 
p_{21} & p_{22} & \cdots & p_{2g} 
\vdots & \vdots & \ddots & \vdots 
p_{m1} & p_{m2} & \cdots & p_{mg}
\end{bmatrix}, \quad p_{ij} = \begin{cases} 1 & \text{if } i^{th} \text{ user } \in j^{th} \text{ group,} \\
0 & \text{otherwise.}
\end{cases}$$

(3.6)

where $\hat{u}_i$ is a row in user matrix $\hat{U}$ (see Eq. 3.1) and $p_{ij}$ is the flag in a membership-indicator matrix $P$, indicating belonging to the group or not. Thus, the dimensions of $K$ (shape) are dependent on how the belonging function is defined.

Examples of aggregations:

- **Preferred Entity** - in order to characterize an entity by its relations to other entities one might assume each user has a *preference* for one entity, and that the remainder entities are *disfavored*. A user’s favorite entity is the one to which he/she mentions the most in posts. We encode this user-entity preference in the membership-indicator matrix $P$ by conditioning $i^{th}$ user $\in j^{th}$ group if $j = \arg \max_j \hat{U}(i, j)$. In this case, $K = [k_{bj}]_{n \times n}$ is the entity–entities characterization (square) matrix. For instance, this approach can
be used to extract information regarding the entities in the context. It can unveil their relationships such as cooperation and rivalries between them.

- **Regions** - on the characterization of regions based on their relationship to entities one can defines regions themselves as the groups, and perceives the “belonging to” function as the “living in” function. In this case, $K = [k_{bj}]_{r \times n}$ is the region–entities matrix, where $r$ is the number of regions. For instance, this approach shows how different regions relate to the same set of entities, and can help to understand the role of neighboring (geographical closeness), common historic background, or any other social-economic indicator in their relationship.

The group-characterization matrix $K$ can be written in a more compact form to improve performance and to simplify its calculation. In Equation 3.7, $P^T P$ gives a diagonal matrix containing the number of users of each group and $P^T \hat{U}$ gives the sum of mentioned scores per entity.

$$ K = \begin{bmatrix} k_1 \\ k_2 \\ \vdots \\ k_g \end{bmatrix} = (P^T P)^{-1} P^T \hat{U}. \quad (3.7) $$

### 3.2.3 Validation

A data analysis project can be used to provide new approaches retrieve known information. In such cases, the efficacy is validated by comparing the results obtained by the new approach and the real known ones (expected), i.e., a direct
measurement of the error. However, even when projects are designed to discover new information, i.e., unveil relationships or statistics hidden within data, we have to validate them. A possible approach is to use deductive reasoning. By showing that a sequence of steps are correct, one might assume the last step is also correct. In a pragmatic approach, if the data is correct and the characterization is correct, then the results provided by a established technique analyzing the characterization should be correct too.

The characterizations proposed here are derived from data, and therefore, they are reliable to represent the data. The question to validate is whether we can use these characterizations to extract knowledge about the data. For instance, we need to check how stable these characterizations are regarding to sample size and to time window variations. We proposed to group users based on their preferred entity; so, we need to investigate at what extent users have well-defined preferences. Then, look for preference changes (i.e., preferred entity mutable), asserting that it is not likely to a preferred entity become non-preferred. There is no correct answer for the questions above, because each problem may have different requirements. However, they should be addressed in order to guarantee the correct interpretation of the results.

Then, we propose a two-level guideline to help addressing the validation of characterizations:

**User Characterization**

1. Types of User – since the datasets are obtained from real social networks, users are expected to post at very different paces, i.e., the distribution of the number of posts per users tend to have a heavy tail, such as in a powerlaw or truncated-powerlaw distributions [26]. Thus, to
evaluate the extent to which distinct types of users behave differently, requiring special treatments. Ideally, all users should be handled in the same way.

2. Rank Stability – Ghoushal and Barabasi proposed a method to evaluate the stability of nodes’ pageranks for different network topologies [52]. They measure the stability by comparing the pagerank values of consecutive nodes (in the rank) after perturbations in the network (random rewiring). Borrowing their ideas, we propose to evaluate the rank stability of the mention received by entities. As defined in Section 3.2.1, the user characterization is a distribution of the frequencies of mentions to a set of entities. To understand the stability of the ranked position of the entities being mentioned, we build a user characterization at each new post cumulatively, from his/her first post ($t = 1$) until the last ($t = T$). Every new post can be seen as a perturbation. Then, for each new perturbation, we calculate the difference between each ranked frequency. At the end, we say the ranking of an entity at rank $m$ is stable if the average difference between its mentioned score and the mentioned score of following rank $\Delta(x^m)$ is greater than their standard deviation $\delta(x^m)$. Formally, the $m$th ranked entity has a stable rank if

$$\delta(x^m) \leq \Delta(x^m), \quad (3.8)$$

$$\Delta(x^m) = \frac{1}{T} \sum_{t=1}^{T} x_t^m - x_t^{m-1}, \quad (3.9)$$
\[ \delta(x^m) = \sqrt{\frac{1}{T-1} \sum_{t=1}^{T} [(x^m_t - x^m_{t-1}) - \Delta(x^m)]^2}, \quad (3.10) \]

where \( x^m_t \) is the mentioned score of the entity at \( m \)th ranked position when characterizing a user considering his/her first \( t \) tweets. The rank stability evaluates the capacity of the metric in providing stable ranked entities.

3. Temporal Fluctuations – stable does not mean immutable, therefore, it is important to check how the \( n \)-stable positions fluctuates along time. Assuming the whole data represents the truth, for each user, we can track the final \( n \)-ranked entities by plotting their cumulative position since the first post. The average behavior can reveal trends and seasonalities, and help to define the length for valid longitudinal investigations over the data.

4. Singularities – depending on the problem, it may be useful to determine the specificity provided by the characterization. For instance, the distance between the first characterization and the last of the same user or among different users.

**Aggregated Characterization**

1. Rank Stability – the same approach as defined for the user characterization. In addition, one may compare the effects of the aggregation. For instance, check if it leads to more stability or not.

2. Sample Size – by definition, the aggregated characterization is a composition of user characterizations under some criteria. As for the user characterization, the aggregation represents frequencies of mention to
entities. A way to evaluate the representativity of the aggregation is to compare the rank of entities for different sample sizes.

3. Temporal Fluctuations – the same approach as defined for the user characterization.

Yet, the validations suggested here are not mandatory to perform the characterizations themselves, but they are tools to ensure characterization results. In other words, in a practical scenario, the validation should be used in a conceptual phase of projects, but probably not on real-time applications.

3.3 Analyzing the Characterizations

In the previous section, we showed how to transform unstructured social data into structured data (characterizations). In this section, the third step of the methodology (see Figure 3.3), we propose approaches to shed light on the underlying processes involving these entities, and to provide means to display or summarize the data.

The relationship among entities can be explored by ranking entities preferences or by identifying unexpected prevalence in some places. Sometimes, the relation-

![Figure 3.3: The analyses phase consist of applying traditional computational methods (e.g., visualization, rank, cluster, and network), using the characterizations as input to extract knowledge from data.](image)
ships need to be more specific or contextualized. Depending on the problem, it might be helpful to identify groups or to define categories. Ultimately, depending on the size of the set of entities evaluated, one might try to understand the structure of their relationship.

### 3.3.1 Ranking

In many contexts, it is natural to compare entities in order to reveal preferences, dislikes, coverage, or market share to cite a few.

Given the characterizations of users and groups, we propose indexes for measuring entities “empathy” and “apathy”. These two indexes form the basis of our entity-ranking strategy to assess the attention entities receive in posts. We define the *empathy index* of entity $j$ as the sum of all attention given to that entity by all users, i.e.:

$$
e_j = \frac{1}{m} \sum_{i=1}^{m} \hat{u}_{ij}. \tag{3.11}$$

To define apathy, we need to apply the preferred entity group characterization concept above mentioned. In contrast to empathy index, the entity’s *apathy index* is all the attention given to an entity by its non-preferred users, which is given by:

$$a_j = \sum_{i=1}^{m} \hat{u}_{ij} (1 - p_{ij}) \sum_{i,j=1}^{m,n} \hat{u}_{ij} (1 - p_{ij}). \tag{3.12}$$

Figure 3.4 shows a visual representation of the similarities and differences between these two proposed indexes as well as the preferred entity group-characterization example.
Figure 3.4: Visual representation of a normalized contingency matrix $\hat{U}$ sorted by preferred entities. It displays: (i) users as rows; (ii) entities as columns; (iii) empathy index $e_j$ aggregates mentioned scores over an entire column (green); (iv) apathy index $a_j$ ignores mentioned scores from preferred users (red); and (v) users with predilection to an entity used to define an entity-entities group-characterization $k_j$ (blue).

### 3.3.2 Highlighting

Once one has regions characterized by entities, a possible analysis is to compare entities’ rank among regions. However, depending on the context, entities’ prevalence can be very unbalanced. For example, the number of supporters from a small-neighbor club versus an international one; or the prevalence of flu against AIDS. Thus, simple ranking entities might not be enough to capture variations on minority groups.

To minimize this problem when comparing regions, instead of assess a region based on any absolute measurement of entity rank (e.g., empathy index), we calculate the Relative Risk [143, 176] of entity $j$ in region $r$ as:

$$RR_{rj} = \frac{e_{in}^j}{e_{out}^j},$$  \hspace{1cm} (3.13)
where $e_j^{in}$ and $e_j^{out}$ are the empathy index for entity $j$ within and outside region $r$, respectively. In this sense, a $RR_{rj} > 1$ indicates an excessive empathy for $j$ in $r$ relative to the overall $j$’s empathy in other regions.

Since the distribution of $\log(RR)$ is approximately normal, entities significantly exceed their expected proportion (highlighted) in a region if $\log(RR) - SE \times 1.96 > 0$, with 95% confidence level [68].

### 3.3.3 Visualizing

Group characterizations are the collective behavior of a population representing preferred entities or regions, and they are encoded as distributions of mentioned scores. Additionally, a visual representation of these distributions by itself allows one to understand roles and interactions among the entities being studied.

In order to get a visual representation of characterizations, we need to map the distribution of mentioned scores in a 2D representation (scatter, bars, etc.). There are several ways to map $k_j$, in particular, we propose the Y-axis to be mentioned score while the X-axis can be:

**Ranked** – a mapping between the entities and the X-axis based on the mentioned score rank. This approach is useful to compare how groups behave. When comparing characterizations, any position on X-axis of two distributions relates to the same rank, not necessarily the same entity. For example, on Figure 3.5 (left), $G3 > G2 > G1$ in the first rank position. In this way, $G1$ looks more similar to $G2$ than to $G3$ in the sense of how (intensity) they share their attention, i.e., $G1$ is closer to $G2$ in all rank positions. Formally,

$$k^r_j : (\forall k_{jx} \forall k_{jy} \in k_j) \ k_{jx} > k_{jy} \rightarrow x < y,$$

(3.14)
Figure 3.5: Different approaches to represent grouped characterizations visually. On left, the ranked approach focusing on how is the behavior, while on right the labeled approach focusing on which entities are receiving attention.

where, $k_j^r$ is the ranked version of $k_j$ with bins sorted by mentions, i.e., the first bin represents the amount of exclusive attention users give to their most-mentioned entity, the second bin represents the attention given to their second entity, and so on.

**Labeled** – a fixed mapping between the entities and the X-axis. This approach is useful to focus on which entities are captured by the relationships being modeled. When comparing characterizations, any position on X-axis of two distributions relates to the same entity. For example, on Figure 3.5 the clear similarity between $G_1$&$G_2$ on the ranked representation (left) is not present on the labeled one (right). Furthermore, $G_1$ is closer to $G_3$ in 75% of the entities ($A, B$, and $D$). In this case, $G_1$ looks more similar to $G_3$ than to $G_2$ since they tend to agree on which entities they give their attention.
3.3.4 Grouping

The visual representation of characterizations exhibits some patterns such as height, length, and curvature, or in a broad sense, shape signatures. As they are distributions, they are already represented as feature vectors, and therefore, they can be used by traditional clustering algorithms in order to reveal common behaviors. As for the visualization, the distributions can be used labeled (unranked) to find groups based on to whom attention is given to or they can be ranked to find groups based on how attention is divided among entities.

To measure shape similarities, distance metrics such as the Euclidean distance might be unsuitable. A better way to compare probability distributions is to use an $f$-divergence function, such as the Hellinger distance [161]. The Hellinger distance is based on the Bhattacharyya coefficient, and it is sometimes called Bhattacharyya distance\(^2\) [69]. Given two histograms $P$ and $Q$, the Hellinger distance $H(P, Q)$ is

$$H(P, Q) = \sqrt{1 - \frac{1}{\sqrt{PQN^2}} \sum_i \sqrt{p(i) \times q(i)}}.$$  (3.15)

It is worthy to note that the Hellinger distance is the Euclidean norm of the difference of the square root of $P$ and $Q$ when they are discrete probability distributions. That is, it can be calculated as

$$H(P, Q) = \frac{1}{\sqrt{2}} \| \sqrt{P} - \sqrt{Q} \|_2.$$  (3.16)

The Hellinger distance is bounded $[0, 1]$, symmetric, and obeys the triangle inequality---

---

\(^2\) Throughout this dissertation, we may use Hellinger distance or Bhattacharyya distance interchangeably.

43
ity. In addition, it is less affected by non-normal distributions, and it measures the similarity of pairs independently (e.g., it does not rely on covariance matrix) [128].

We propose to investigate the extent of similarity among entities by using the Agglomerative Clustering algorithm from the scikit-learn package [121] with the Hellinger distance as the affinity metric. The hierarchical clusters can provide extra information to the specificity of an entity distribution visualization (Section 3.3.3). It is also important to test the Cophenetic Correlation Coefficient [145] to determine the best linkage method to be used.

3.3.5 Connecting

As we showed, one can explore the aggregated characterizations at different granularities. We can focus on specific aggregations, and use visualizations to understand the relationships intra-entities (Section 3.3.3); or we can use clustering to explore the relationships inter-entities (Section 3.3.4). However, when the number of entities are large, we can explore a higher level granularity using networks to understand the overall structure of the relationships and any other emerging effect of that.

There are several forms to map the characterizations to a network, but we propose a straightforward approach, i.e., to use the group-characterization matrix $K$ (Equation 3.7) as the adjacency matrix of the network. Thus, the network will reflect the aggregated function chosen. For instance, using the preferred entity approach will raise a directed-weighted-asymmetric network where the nodes are the entities, and the edges represent the strength of engagement of source-entity-labeled users on the target-entity. On the other hand, using the regions approach will raise a bipartite network of entities and regions. Once we have a network,
we can explore Network Science analyses, such as to extract network properties, to identify central nodes, to test the network resilience, to explore its capacity to diffuse information, or to detect communities.

3.4 Experience

This last section of the chapter and step of the proposed methodology, as illustrated in Figure 3.6, is more abstract and more problem dependent. While the previous sections showed pragmatically how to move from unstructured data to a standardized form able to be used by several established techniques, this last step is motivational. After the exploratory analysis over the data and with new information available, the scientific process should continue. The information generated so far is interesting, but at the same time it is shallow. One must keep in mind how to use it to answer deeper questions.

![Figure 3.6: The last phase is experience, i.e., to use the prior knowledge, already generated by the process, to delve further knowledge.](image-url)
Chapter 4

Learning from Social Media Data

In this chapter, we instantiate the proposed methodology in three distinct domains: football [114], language [115], and organ donation [113]. In all applications, we (i) defined sets of entities and how to account for mentions; (ii) applied different group-characterizations approaches; and (iii) performed several analyses. Figure 4.1 illustrates this process of learning from social media data by transforming unstructured data into information.

Figure 4.1: Instantiating the methodology. After choosing the context, define the set of entities and the research questions. Then, plan adequately which group-characterizations and analyses will help you to find the answers.
4.1 Characterizing Football Supporters

Football (aka soccer) is by far the most popular sport in the world with an estimate of 3.5 billion fans worldwide\(^1\). The popularity of the sport leads to several stories (some perhaps anecdotal) about supporters behaviors and classic rivalries. Little, however, has been done to characterize/profile our behaviors as football supporters using social media data, and to use our behavior as an aggregate measure of 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 application, we use techniques from data science, comprising our methodology (see Chapter 3), to characterize football supporters according to their activity on Twitter. We show that it is possible to characterize clubs according to the behavior of their supporters, and that specific signatures repeat themselves across different clubs and in different countries. Our results were published in the 2016 IEEE/WIC/ACM International Conference on Web Intelligence [114].

4.1.1 Motivation

Football is by far the most popular sport in the world with an estimate of 40\% more supporters worldwide than cricket, the second most popular sport which has 2.5 billion 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. The popularity of the

\(^1\)http://www.topendsports.com/world/lists/popular-sport/fans.htm
sport leads to several stories about supporters behaviors and 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 our 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) [61], or even by sponsors who sometimes need to understand the behavior of supporters and their clubs [62, 110].

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. We have identified signatures of how supporters behave and then used the signatures to classify clubs in Brazil and in the UK\(^2\) 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 or are a consequence of social pressures or a phenomenon that is completely independent.

\(^{2}\)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”.

48
4.1.2 Experiment Settings

4.1.2.1 Methodology Instantiation

In this section, we detail how the proposed methodology was instantiated in the football supporters context.

- **Data** – two Twitter datasets and several traditional ranks for number of supporters and most-hated clubs.

- **Entity** – a football club.

- **Mention** – club’s Twitter official account being mentioned (@) in the tweet text.

- **Set of Entities** – 20 Brazilian clubs and 20 English clubs.

- **Aggregation Method** – preferred entity, thus the group-characterization represents a club characterization.

- **Analyses** – ranking, visualizing, and grouping (labeled and ranked).

4.1.2.2 Characterizing Supporters and Clubs

As defined on Section 3.2, 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 extract these mentions about clubs from large datasets of tweets sent during competition seasons 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. In this context, 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 mentions to club $j$ in the tweets by user $i$ (see Section 3.2.1).

To characterize clubs, we further assume that a user supports one club and that the remainder clubs are opponents; Figure 4.2 evidences the plausibility of this assumption by showing the distribution of mentioned scores clearly skewed towards a single preferred club. A user’s favorite club is the one to which the user mentions the most in tweets. We encode this user-club preference in the membership-indicator matrix $P$ using the preferred entity aggregation approach (see Section 3.2.2). Once the user’s preferred club is known, the other clubs mentioned in tweets by that user are considered to be opponent clubs. Therefore, the group-characterization matrix $K$ represents the club-characterization matrix.

Once we have clubs characterized, and since their nature is to compete to each other, we may ask:

1. To what degree can clubs be ranked? How does the ranking relate to tradi-

![Figure 4.2: Distribution of non-zero mentioned scores ($\hat{u}_{ij}$) for all users (supporters) characterized in the Brazilian (in blue) and in the English (in red) datasets. The peak closer to 1 indicates users tend to mention a specific club much more than all others. The filled area highlight which distribution is greater.](image)
tional existing ranks?

2. Can pairwise rivalries be identified?

3. To what degree can clubs be categorized based on number of rivals? And based on rivalry similarity?

4.1.2.3 Datasets

For this application, we used two football datasets collected from Twitter: one representing the 2014/2015 English Premier League (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 4.1.

Table 4.2 presents some statistics about these two datasets. If we assume tweets and users are 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 EPL 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 Figure 4.3(c)). Overall, users on BSA dataset mention more clubs than those on EPL, 1.98 and 1.47 respectively (see Figure 4.3(b)). This behavior reflects on the normalized contingency matrix distribution, where BSA supporters present higher density in the lower-score region (see Figure 4.2). Finally, we analyzed the distribution of the number of tweets per user in both leagues in Figure 4.3(a). We found that both datasets fit better on log-normal distributions than power-law or exponential distributions, an expected phenomena over social
Table 4.1: Twitter official accounts from clubs used in football supporters characterization.

<table>
<thead>
<tr>
<th>Brazilian “Série A” (BSA)</th>
<th>English Premier League (EPL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Club</td>
<td>Account</td>
</tr>
<tr>
<td>A. Mineiro</td>
<td>@atletico</td>
</tr>
<tr>
<td>A. Paranaense</td>
<td>@atleticopr</td>
</tr>
<tr>
<td>Avai</td>
<td>@avaifc</td>
</tr>
<tr>
<td>Chapecoense</td>
<td>@ChapecoenseReal</td>
</tr>
<tr>
<td>Corinthians</td>
<td>@Corinthians</td>
</tr>
<tr>
<td>Coritiba</td>
<td>@coritiba</td>
</tr>
<tr>
<td>Cruzeiro</td>
<td>@Cruzeiro</td>
</tr>
<tr>
<td>Figueirense</td>
<td>@FigueirenseFC</td>
</tr>
<tr>
<td>Flamengo</td>
<td>@Flamengo</td>
</tr>
<tr>
<td>Fluminense</td>
<td>@FluminenseFC</td>
</tr>
<tr>
<td>Goias</td>
<td>@goiasec_oficial</td>
</tr>
<tr>
<td>Gremio</td>
<td>@gremiooficial</td>
</tr>
<tr>
<td>Internacional</td>
<td>@SC Internacional</td>
</tr>
<tr>
<td>Joinville</td>
<td>@jec_online</td>
</tr>
<tr>
<td>Palmeiras</td>
<td>@SEPalmeiras</td>
</tr>
<tr>
<td>Ponte Preta</td>
<td>@aapp_oficial</td>
</tr>
<tr>
<td>Santos</td>
<td>@SantosFC</td>
</tr>
<tr>
<td>Sao Paulo</td>
<td>@SaoPauloFC</td>
</tr>
<tr>
<td>Sport</td>
<td>@sportrecife</td>
</tr>
<tr>
<td>Vasco</td>
<td>@crvascodagama</td>
</tr>
</tbody>
</table>

network observations.

4.1.3 Results

4.1.3.1 Ranking

We can interpret the empathy\(e_j\) and apathy\(a_j\) indexes, proposed on Section 3.3.1, in the context of football supporters as a proxy to clubs popularity and opposition, respectively. We assumed that a user supports the club he/she most mention. In addition to the visual confirmation provided in Figure 4.2, to check the plausibility of this assumption we ranked clubs according to our proposed indexes of empathy.
Table 4.2: Some statistics about the Brazilian and English football datasets.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>BSA</th>
<th>EPL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start Data Collection</td>
<td>25/May/15</td>
<td>07/Feb/15</td>
</tr>
<tr>
<td>Finish Data Collection</td>
<td>08/Dec/15</td>
<td>07/May/15</td>
</tr>
<tr>
<td>Number of Days</td>
<td>198</td>
<td>89</td>
</tr>
<tr>
<td>Number of Matches</td>
<td>35 / 38</td>
<td>12 / 38</td>
</tr>
<tr>
<td>Season Coverage Interval</td>
<td>8%–100%</td>
<td>60%–92%</td>
</tr>
<tr>
<td>Tweets with Mentions</td>
<td>7,578,456</td>
<td>4,920,764</td>
</tr>
<tr>
<td>Users with Mentions</td>
<td>626,208</td>
<td>1,151,702</td>
</tr>
<tr>
<td>Percentage of Balanced Users</td>
<td>6.13%</td>
<td>7.71%</td>
</tr>
<tr>
<td>Avg. Tweets / User</td>
<td>12.10</td>
<td>4.27</td>
</tr>
<tr>
<td>Avg. Tweets / Day</td>
<td>38.469</td>
<td>55.289</td>
</tr>
<tr>
<td>Avg. Tweets / User / Day</td>
<td>0.061</td>
<td>0.048</td>
</tr>
<tr>
<td>Avg. Mentions to Clubs / Tweet</td>
<td>1.10</td>
<td>1.07</td>
</tr>
<tr>
<td>Avg. Clubs Mentioned / User</td>
<td>1.98</td>
<td>1.47</td>
</tr>
</tbody>
</table>

(as popularity, Equation 3.11) and apathy (as opposition, Equation 3.12). Then, we compared our ranks against ranks of supporters, most-hated clubs, and current season ranks using Spearman correlation [84].

4.1.3.1.1 BSA Ranks

For the Brazilian dataset, we found four rankings regarding the amount of supporters, and one about most hated club. 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 [46]. The margin of error was 1.5% at 95% confidence level. This is the only Brazilian poll to consider the most-hated club we had access to.

Datafolha (DF14) – 4,337 people, in June of 2014, were interviewed by this poll from Datafolha institute [70]. Error margin was 2% at 95% confidence level.

Ibope (I14) – 7,005 people, in 2014, participated in the poll requested by website
Figure 4.3: Three plots describing two Twitter datasets: Brazilian League (in blue) and Premier League (in red). (a) Distributions of number of tweets per user: real data in solid lines, log-normal in dotted lines, and power-law in dashed lines; (b) Number of clubs mentioned per user, ranging from 1 to 20 (max per league); (c) Number of clubs mentioned per tweet.

Lance! to IBOPE research institute [56]. 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 [30]. 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 4.3 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.
Table 4.3: BSA – proportions, number of users, and rank positions for clubs based on (left) popularity $e_j$ and on (right) opposition $a_j$, compared against Datafolha (DF14), Ibope (I14), and Paraná (P13 & P16) great supporters and most hated club.

<table>
<thead>
<tr>
<th>Club(Position)</th>
<th>$e_j$</th>
<th>#Users</th>
<th>Rank</th>
<th>DF14</th>
<th>I14</th>
<th>P13</th>
<th>P16</th>
<th>P16 Pos*</th>
<th>Correlation with $e_j$ rank</th>
<th>Club</th>
<th>$a_j$</th>
<th>#Users</th>
<th>Rank</th>
<th>P16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flamengo 17.32 108450 1 1 1 1 1 12</td>
<td>Corinthians 13.51 10894 1 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corinthians 16.73 104746 2 2 2 2 2 1</td>
<td>Flamengo 10.69 8619 2 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sao Paulo 10.42 65237 3 3 3 3 3 4</td>
<td>Sao Paulo 7.62 6147 3 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Internacional 8.40 52593 4 7 9 9 10 5</td>
<td>Fluminense 7.39 5958 4 -</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cruzeiro 6.27 39287 5 7 7 6 6 8</td>
<td>Internacional 7.38 5953 5 8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fluminense 5.93 37132 6 10 11 14 13 13</td>
<td>Vasco 7.23 5829 6 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Palmeiras 5.79 36237 7 4 4 4 4 9</td>
<td>A. Mineiro 5.74 4632 7 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gremio 5.71 35760 8 6 8 7 7 3</td>
<td>Chapecoense 5.66 4561 8 -</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vasco 5.02 3409 9 5 5 5 5 18</td>
<td>Cruzeiro 5.62 4529 9 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Santos 4.81 30122 10 7 10 8 8 7</td>
<td>Palmeiras 5.53 4462 10 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A. Mineiro 4.22 26415 11 10 6 10 9 2</td>
<td>Gremio 5.48 4420 11 9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sport 2.43 15224 12 12 15 11 14 6</td>
<td>Santos 5.08 4096 12 -</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chapecoense 2.20 13771 13 - - - - 14</td>
<td>Sport 2.62 2110 13 -</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coritiba 1.15 7207 14 - - 19 - - 15</td>
<td>Figueirense 2.12 1712 14 -</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A. Paranaense 1.06 6616 15 - 15 17 - - 10</td>
<td>Coritiba 2.11 1698 15 -</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Figueirense 0.89 5575 16 - - - - - 16</td>
<td>A. Paranaense 1.67 1344 16 -</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avai 0.79 4938 17 - - - - - 17</td>
<td>Avai 1.67 1342 16 -</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Joinville 0.34 2123 18 - - - - - 20</td>
<td>Joinville 1.10 884 18 -</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Goias 0.29 1830 19 - - 20 - - 19</td>
<td>Goias 0.97 783 19 -</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ponte Preta 0.24 1526 20 - - - - - 11</td>
<td>Ponte Preta 0.83 670 20 -</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Popularity and opposition indexes results in % with confidence intervals $e_j \pm 0.25\%$ and $a_j \pm 0.62\%$ at 99% confidence level. Significant codes for $p$ values: $0 \‡ .001 \† .01 \* .05$.

The popularity index ($e_j$) 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 ($a_j$) did not correlated as high as the popularity one, $a_j$ 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.

### 4.1.3.1.2 EPL Ranks

The comparative ranks for the English clubs were created by SportsMail on 2015 [65]. 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 (empathy). The descriptive text of 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.”

In addition, we used the survey performed by the Mirror newspaper [35] 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 4.4 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 ($r = .71$, $p < .01$).

Finally, the correlation with the current season position showed the worst results for both datasets suggesting $e_j$ captures supporters’ long-term preferences. Additional club rankings, such as by the number of user, tweets, and preferred users can be found in appendixes B.2 and B.3.
Table 4.4: EPL – proportions, number of users, and rank positions for clubs based on (left) popularity $e_j$ and on (right) opposition $a_j$, compared against objective ranks from SportsMail as great supporters and Mirror15 as most hated club.

<table>
<thead>
<tr>
<th>Club</th>
<th>$e_j$ * #Users</th>
<th>Rank</th>
<th>Crowds</th>
<th>Fanbase</th>
<th>Total</th>
<th>Pos*</th>
<th>Club</th>
<th>$a_j$ * #Users</th>
<th>Rank</th>
<th>M15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Man Utd</td>
<td>28.39 326931</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>Man Utd</td>
<td>19.15 27150</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Arsenal</td>
<td>17.78 204729</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>Liverpool</td>
<td>17.23 24419</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Chelsea</td>
<td>17.81 205076</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>Chelsea</td>
<td>13.31 18870</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Liverpool</td>
<td>14.10 162348</td>
<td>4</td>
<td>10</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>Arsenal</td>
<td>13.24 18770</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Man City</td>
<td>6.09 70099</td>
<td>5</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>Man City</td>
<td>8.99 12744</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Spurs</td>
<td>2.91 33481</td>
<td>6</td>
<td>8</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>Aston Villa</td>
<td>3.57 5066</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>Aston Villa</td>
<td>2.01 23170</td>
<td>7</td>
<td>9</td>
<td>7</td>
<td>8</td>
<td>17</td>
<td>Spurs</td>
<td>3.56 5041</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>Everton</td>
<td>1.80 20768</td>
<td>8</td>
<td>6</td>
<td>8</td>
<td>6</td>
<td>11</td>
<td>Everton</td>
<td>3.22 4570</td>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td>Newcastle</td>
<td>1.45 16604</td>
<td>9</td>
<td>7</td>
<td>9</td>
<td>9</td>
<td>15</td>
<td>Newcastle</td>
<td>2.83 4006</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>West Ham</td>
<td>1.24 14312</td>
<td>10</td>
<td>22</td>
<td>10</td>
<td>12</td>
<td>12</td>
<td>Swansea</td>
<td>2.57 3644</td>
<td>10</td>
<td>18</td>
</tr>
<tr>
<td>Sunderland</td>
<td>0.95 10931</td>
<td>11</td>
<td>5</td>
<td>13</td>
<td>10</td>
<td>16</td>
<td>Sunderland</td>
<td>1.59 2249</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>Southampton</td>
<td>0.74 8487</td>
<td>12</td>
<td>34</td>
<td>11</td>
<td>15</td>
<td>7</td>
<td>Crystal Palace</td>
<td>1.54 2189</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>QPR</td>
<td>0.71 8227</td>
<td>13</td>
<td>36</td>
<td>16</td>
<td>31</td>
<td>20</td>
<td>West Ham</td>
<td>1.46 2066</td>
<td>13</td>
<td>9</td>
</tr>
<tr>
<td>Swansea</td>
<td>0.69 7930</td>
<td>14</td>
<td>39</td>
<td>12</td>
<td>34</td>
<td>8</td>
<td>QPR</td>
<td>1.42 2017</td>
<td>14</td>
<td>-</td>
</tr>
<tr>
<td>Crystal Palace</td>
<td>0.71 8181</td>
<td>15</td>
<td>18</td>
<td>19</td>
<td>30</td>
<td>10</td>
<td>West Brom</td>
<td>1.18 1667</td>
<td>15</td>
<td>13</td>
</tr>
<tr>
<td>Leicester</td>
<td>0.64 7320</td>
<td>16</td>
<td>21</td>
<td>22</td>
<td>21</td>
<td>14</td>
<td>Southampton</td>
<td>1.11 1576</td>
<td>16</td>
<td>19</td>
</tr>
<tr>
<td>West Brom</td>
<td>0.63 7296</td>
<td>17</td>
<td>11</td>
<td>20</td>
<td>11</td>
<td>13</td>
<td>Leicester</td>
<td>1.06 1508</td>
<td>17</td>
<td>13</td>
</tr>
<tr>
<td>Burnley</td>
<td>0.55 6386</td>
<td>18</td>
<td>27</td>
<td>31</td>
<td>26</td>
<td>19</td>
<td>Hull</td>
<td>0.99 1400</td>
<td>18</td>
<td>-</td>
</tr>
<tr>
<td>Hull</td>
<td>0.39 4543</td>
<td>19</td>
<td>19</td>
<td>14</td>
<td>32</td>
<td>18</td>
<td>Burnley</td>
<td>0.99 1398</td>
<td>18</td>
<td>-</td>
</tr>
<tr>
<td>Stoke</td>
<td>0.42 4783</td>
<td>20</td>
<td>17</td>
<td>17</td>
<td>16</td>
<td>9</td>
<td>Stoke</td>
<td>0.98 1395</td>
<td>20</td>
<td>6</td>
</tr>
</tbody>
</table>

* Popularity and opposition indexes results in % with confidence intervals $p_j \pm 0.22\%$ and $o_j \pm 0.54\%$ at 99% confidence level. Significance codes for $p$ values:
0‡ \( 001 \) † \( 01 \) ‡ \( 05 \).

#4.1.3.2 Clubs as Distributions

Since we already discussed the applicability of $e_j$ and $a_j$, now let us explore club characterizations $k_j$. Indeed, $k_j$ allow us to understand 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, we can use the visual representation proposed on Section 3.3.3.

Figures 4.4(a) and 4.5(a) show BSA and EPL clubs ranked distribution (see Equation 3.14), highlighting the top three rivals of each club. This straightforward representation is able to capture famous derbies such as: Atlético Mineiro vs. Cruzeiro, Flamengo vs. Vasco, and Grêmio vs. Internacional, in Brazil; Newcastle
Sunderland, Manchester United vs. Manchester 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 main rival of Flamengo; yet, Liverpool is the rival of Everton, but Manchester 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 their opponents.

Moreover, club distributions show patterns such as height, length, and curvature which can be seen as shape signatures, and in the future they could be used to do a even more accurate classification of supporters considering many clubs around the world. We argued in the motivation of this application that this work could lead to better safety in stadiums; the support signature for clubs may be an indicative 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.

4.1.3.3 Clustering Clubs

Grouping clubs by their ranked signature can reveal similarities on how their fans support, or to whom they support depending whether distributions are ranked or labeled, respectively.

Figures 4.4(b) and 4.5(b) show the similarity matrix based when distributions are ranked. The dendrograms show the hierarchical clusters, and a possible interpretation for them is how competitive the clubs are, i.e how their supporters perceive rivalries among them.
Figure 4.4: Analysis for the 2015 “Série A” Brazilian league dataset. On (a), clubs as distributions and on (b) and (c) clustering approaches.

Figures 4.4(a) and 4.5(a) can aid understanding clusters’ rules in figures 4.4(b) and 4.5(b). 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 (11th) and Goias (19th), but 5 in UK – Southampton (7th), Crystal Palace (10th), Stoke (9th), QPR (20th), and Swansea (8th). So, in terms of rivalry, we have 18 and 15 clubs
Figure 4.5: Analysis for the 14/15 Premier league dataset. On (a), clubs as distributions and on (b) and (c) clustering approaches.

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 (16th), Avai (17th), and Joinville (20th), and 4 clubs in UK – Man. City (2nd), Arsenal (3rd), Man. United (4th), and Chelsea (1st). 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.

Figures 4.4(c) and 4.5(c), on the other hand, present an entire different cluster compositions. They are the results of grouping clubs using a labeled representation, i.e., focusing on the rivalry similarities. This time, clusters groups clubs based on geographical proximity; for instance, in the Brazilian dataset clubs are mostly grouped by states. This effect is stronger in the BSA because Brazil has additional leagues for each state, perhaps nurturing even more rivalry.

4.1.4 Discussion

In this first application of the methodology proposed on Chapter 3, 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 aggregated characterization of several supporters (their vectors) leads to clubs characterization.

We evaluated two measurements for clubs: popularity and opposition. 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 1.7 million 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.

There are still many questions to be addressed by this work. We list below 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.
4.2 Characterizing Multilingual Users

Online social networks (e.g., Twitter) offer an open platform for people to interact and to connect without restrictions of language usage or geographic borders. Because of their pervasiveness, online social networks provide data and become real-time sensors of society. This application looks at Twitter to reveal the hidden relationship of languages that stems from users’ language preference for writing their tweets. We show that the language relationships are dependent of place by comparing 12 large-scale datasets with different locality levels. For instance, the secondary language of French speakers in Canada is different from French speakers in France. We used network science and clustering techniques to find that languages groups are more driven by spatial than syntactic proximity. The characterization of language relationships is key to understanding information spread in social media and the detection of cultural shifts. Our results were published in the 14th IEEE International Conference on Ubiquitous Intelligence and Computing (UIC 2017) [115].

4.2.1 Motivation

The effect of globalization over the past few years has been observed in various domains of our lives, including trading, immigration, education, and culture [29, 122, 64, 151]. Due to the connectedness of society, information (fads, trends, etc.) tends to be transmitted through people’s social networks. This effect on society has been extensively discussed in the literature [81, 51, 57, 155, 156]. Yet, the impact of language to information spread has received very little attention. Today, a handful of languages have become globally popular; the popularization of TV in the late 50s
and early 60s, as well as the current explosion in the use of social media, have all contributed to the popularity of certain languages. Although historic relationships among the languages are useful, understanding the significance of the language relationships from population preferences can lead to the identification of possible culture shifts.

Languages may be organized hierarchically according to historical relationships leading to language family trees; examples of a few popular language families are: Afro-Asiatic, Dravidian, Indo-European, Tai-Kadai, and Uralic. In a language tree, the closer the languages are from each other, the more similar they tend to be syntactically. Figure 4.6 depicts an illustrative example mainly about the Indo-
European branch. Languages are not static, but evolve with society; for instance, Greek, Arabic, and Latin used to be popular, but today English is considered the *de facto* global language. If one wants to study language relationships in today’s world, the analyses have to consider social media, given its wide use. Social media has become the standard form of communication for the younger generation, as it provides an easy way for them to express their opinions [49].

In spite of several works related to the language of users in Twitter [125, 100, 130, 133, 116], the research community did not pay enough attention to the importance of *language relationships* generated from the user preference. Our work explores the characterization of languages as an emergent effect of individual online behavior.

First, we characterize users based on the frequency of the languages used in their tweets. Then, we aggregate users based on their most used language. We use unsupervised machine learning algorithms and network science to understand the extent to which languages group together due to their origins (family trees) or other factors, such as geographical proximity. We use 12 different Twitter datasets (Table 4.5) to understand and to capture language singularities.

This work reveals the structure of languages on Twitter, and provides a window into how these languages are related to one another in this social media platform. Moreover, we demonstrated how a language relationships can differ from place to place, and how to obtain a global characterization. The characterization we provide can be used to improve target campaigns, marketing strategies, or social network interventions. Potentially, it can be used as a tool to identify unexpected migrations. Finally, we provide some insightful visualizations on the structure of language relationships.
4.2.2 Experiment Settings

4.2.2.1 Methodology Instantiation

In this section, we detail how the proposed methodology was instantiated in the language characterization context.

- **Data** – twelve Twitter datasets.

- **Entity** – a language.

- **Mention** – a BCP 47 language identifier\(^3\) corresponding to the automatically detected language of the tweet text by Twitter in the field `lang`.

- **Set of Entities** – languages identified in each dataset (refer Table 4.5), ranging from 43 to 66 languages (see Figure 4.9 for an example of the detected languages).

- **Aggregation Method** – preferred entity, with the group-characterization representing language characterizations.

- **Analyses** – ranking, connecting, visualizing, and labeled grouping.

4.2.2.2 Characterizing Users and Languages

As defined on Section 3.2, we begin by characterizing users as the frequency of languages they use on their posts (tweets), i.e., we characterize multilingualism users. In this context, since our *entities* are languages, a *mention* to a language is captured by the result of a function, i.e., a language detector algorithm. Although

\[^3\text{https://tools.ietf.org/html/bcp47}\]

68
a tweet can have words from different languages, in our application we only consider tweets automatically detected by Twitter as single-language [118, 153]. That is, those for any reason language-unidentified tweets are discarded from our experiments. Thus, the set of entities is determined by the language detector algorithm.

Let \( \mathcal{T} = \{\tau_1, \ldots, \tau_m\} \) be the set of posts sent by \( m \) users in \( n \) languages, where \( \tau_i \) are the posts sent by user \( i \). As before, we use this data to calculate a contingency table of frequencies of languages per user. The language mentioned score \( \hat{u}_{ij} \) here is calculated using simplified version of Equation 3.3 because we cannot have multiple languages mentioned in a tweet. It is the number of tweets from user \( i \) in language \( j \) divided by the total number of his/her tweets and it is given by

\[
\hat{u}_{ij} = \frac{1}{|\tau_i|} \sum_{t \in \tau_i} \delta(t_\ell, j),
\]

where \( \delta(t_\ell, j) = 1 \) if the tweet language \( t_\ell = j \).

As in the football characterization, we also use the preferred entity aggregation approach (see Section 3.2.2) to characterize languages. That is, we define languages based on users, which in turn, are also defined by languages. So, to characterize a language by its relations to other languages, one might assume each user has a preferred language. A user’s preferred language is the one he/she tweets the most. Therefore, the group-characterization matrix \( K = [k_{jw}]_{n \times n} \) represents the language characterization matrix, where the \( j^{th} \) row is the mean probabilities of all users whose preferred language is \( j \). Consequently, a language is characterized relative to all languages as a distribution of probabilities, i.e., it encompasses the average behavior of its preferred “speakers” (users). It captures the willingness of
the flow of information from one language to the other. In addition, despite not using the regions aggregation as proposed in Section 3.2.2 directly, we mock its expected behavior by using bounded datasets.

Having a model to characterize languages based on their online usage, we may ask:

1. Is there any difference from a language spoken in different regions? For instance, the French speakers on Canada to the ones in France?

2. To what degree can relationships among languages be identified? Is English the real world language?

3. How languages spread worldwide? Is language syntactic similarity the best explanation?

4. Can social media be used to sense languages and to detect cultural shifts in a region?

4.2.2.3 Datasets

The datasets used here vary in the collection process (common terms or bounding boxes), yielding different levels of locality. The datasets that were collected by tracking terms are event-based and, therefore, global; it is instructive to note that their globalization is given not only by the boundless (lake of) restriction whereabouts the post was generated, but from the global appealing of the selected events (e.g., Olympics games and G20 meetings). The other datasets, collected using bounding boxes, are constrained to geographical boundaries, and consequently, they are limited to cities and countries within the box.
Table 4.5: Descriptive statistics of 12 datasets identified by their locality level – global (–red), country(†–green), or city(⋆–blue); the number of tweets and users; the total number of languages |L| and languages used by at least 10 users |L|⁺; the percentage of monolingual users; the average number of languages used by multilingual users; and the period of collection.

| Dataset                                      | Tweets     | Users     | |L| | |L|⁺ | Mono. %/User | From       | To         |
|----------------------------------------------|------------|-----------|-----|-----|-----|----------------|------------|------------|
| 2016 Olympic Games                          | 18,048,522 | 6,506,634 | 61  | 55  | 93% | 2.17           | 08/01/16   | 08/24/16   |
| G20                                          | 10,610,653 | 2,694,784 | 60  | 50  | 93% | 2.28           | 08/24/14   | 09/29/14   |
| 2015 Women’s World Cup & America Cup         | 10,026,573 | 2,704,898 | 62  | 48  | 90% | 2.38           | 06/16/15   | 07/13/15   |
| 2014 FIFA World Cup                          | 50,476,375 | 9,235,153 | 64  | 48  | 73% | 3.10           | 06/12/14   | 07/13/14   |
| 2016 UEFA Euro                              | 36,456,419 | 5,413,895 | 60  | 48  | 79% | 2.78           | 06/10/16   | 07/19/16   |
| The United Kingdom                          | 30,373,072 | 3,069,664 | 65  | 51  | 84% | 2.71           | 02/07/15   | 05/07/15   |
| South America                               | 334,337,096| 2,743,842 | 66  | 44  | 56% | 5.71           | 04/23/15   | 12/08/15   |
| New York City                               | 1,925,831  | 130,368   | 55  | 38  | 80% | 2.92           | 08/29/14   | 09/29/14   |
| Paris                                       | 434,969    | 30,324    | 46  | 34  | 64% | 3.18           | 03/09/15   | 04/03/15   |
| San Francisco                               | 717,555    | 62,989    | 49  | 33  | 82% | 2.77           | 03/05/15   | 05/05/15   |
| Tokyo                                       | 2,153,586  | 147,140   | 57  | 34  | 92% | 2.55           | 03/05/15   | 05/05/15   |
| Hong Kong                                   | 96,302     | 10,682    | 43  | 23  | 61% | 2.93           | 03/05/15   | 05/06/15   |

Table 4.5 shows some statistics of the datasets. Data was gathered within a 3-year period. The datasets vary in number of tweets (0.1–334 million) and users (0.01–9 million). The number of languages used by multilingual users is quite stable among all datasets. However, the number of languages used by at least 10 users |L⁺| on global-level datasets is larger than on city-level, suggesting distinct features among them. More details about the collection process of datasets on Appendix C.

### 4.2.3 Results

#### 4.2.3.1 Locality Effect

First, we investigate at which extent a region may influence how/which languages are written/spoken. Different places are formed by an amalgam of different cultures. Therefore, if regions play a significant role, it is expected that the languages
characterized by using distinct datasets, such as the ones presented on Table 4.5, present different levels of similarities. Moreover, the results and conclusions derived from these results should embed the peculiarities of different places.

A simple method to compare distinct language datasets was proposed by Ronen et al. [130]. That is, to compare the number of speakers of each language in different datasets, and also comparing the number of speakers of pair of languages. The latter comparison is important because it adds the relationship facet we are trying to investigate. We can use the empathy index (see Equation 3.11) to estimate the number of users of a language, i.e., the sum of the partial contribution of all users. Figures 4.7 and 4.8 compare the correlation of the number of users using each language individually as well as those using pairs of languages (e.g., the number of users tweeting in Japanese and Portuguese). Figure 4.9 shows the number of users per language in the Olympics16 global dataset. For other datasets, please refer to Appendix B.4.

The comparison examples using 3 datasets in Figure 4.7 and the overall evaluation in Figure 4.8 suggest datasets tend to be more similar as their locality level decreases. In other words, the relationship between pairs of languages in global datasets tend to be more similar than when compared against the relationship in a city-level dataset. Hence, global datasets are more adequate to describe languages while city ones should be used to discover language singularities within regions. Since this application aims at unveiling the relationships of languages, we focus on the analysis on the most global dataset – Olympics16. Unless explicitly mentioned otherwise, all figures are based on this dataset.
Figure 4.7: Comparing the correlations between pairs of datasets at different levels of locality: global (Olympics16–red), countries (The United Kingdom–green), and city (Tokyo–blue). The correlations are higher for the pairs with the global dataset than for the pairs with the city one. The top row shows the correlation between the number of users tweeting in a language, i.e., each point is a specific language; while the bottom row shows the correlation of users with tweets in pairs of languages, e.g., the number of users using English and Japanese. Comparison between datasets: Olympics16 and The United Kingdom (a and d); Olympics16 and Tokyo (b and e); and The United Kingdom and Tokyo (c and f).

4.2.3.2 Language Network

As proposed on Section 3.3.4, the group-characterization matrix (Equation 3.5), in this context—the language characterization matrix $K$, can be used as an adjacency matrix to build a language network (LN). When $K$ is generated from a global dataset, such as the Olympics16, there is a global LN. The LNs are weighted-directed networks, where nodes are languages, and edge-weights represent the proportion of users interacting in both the source and target languages. For instance, a link of 0.02 from Dutch to English means Dutch users write in English in 2% of their social interactions. The self-edges are ignored in the LN since our interest lie in understanding the relationships among languages; nevertheless, the self-edges

73
Figure 4.8: Measuring the locality effect of datasets on language relationships. Each boxplot contains the Pearson correlations for 11 pairs of datasets. From Hong Kong (top) to Olympics16 (bottom), datasets vary from more local to more global level as depicted by the following special marks: city-level (*—blue), country-level (†—green), and global-level (◦—red). On average, global datasets presents higher correlation than city datasets. On the left, correlations between users tweeting in one language, while on the right, the correlations of users tweeting in pairs of language. The top blocks show individual dataset comparisons while bottom blocks show inter and intra-levels comparisons.

are important while calculating network edge’s weights since they encapsulate the notion of mono- and multilingual users of a particular language.

Table 4.6 presents the top 10 languages based on a few centralities. The \textit{in-degree} represents the diversity of the neighborhood of a language (i.e., how many languages have at least one user who also uses the language in question), while the \textit{weighted in-degree} of a language captures the volume of speakers for each of these connections captured by the in-degree. For instance, Italian is among the top 10
languages with high \textit{in-degree}, while Russian is among the top 10 languages with high \textit{weighted in-degree}; therefore, people from different cultures use more Italian than Russian, but the fewer cultures who use Russian, do it much more frequently. The \textit{out-degree} of a language demonstrates the \textit{multilingualism} of the users of the language, or the tendency of users to connect to others with distinct preferred languages. In the LN, language relationships are asymmetric; the in-degree and out-degree results differ. For instance, although Italian\textsuperscript{4} users interact in several languages, the percentage of Serbian\textsuperscript{4} users who interact in languages other than Serbian is greater than Italians interacting in languages other than Italian. Thus, while in-degree can be seen as a measurement of language popularity, the out-degree can be seen as a plurality indicator of users of a particular language. Finally, the \textit{eigenvector} is an influential centrality (diversity and volume) since it considers the structure of the network.

Figure 4.10(a) shows the language network. The sizes of nodes represent the weighted in-degree, and the colors are the communities according to Blondel’s algorithm [9]. We used language community colors to set countries’ color in the world map on Figure 4.10(b). Each country assumes the color of its major lan-

\textsuperscript{4}In this context, Italian, Serbian, Dutch, etc. refer to the preferred language rather than nationality.
Table 4.6: Top ten languages ranked in the LN based on in/out degree and weighted degree, and eigenvector centralities.

<table>
<thead>
<tr>
<th>Rank</th>
<th>In-Degree</th>
<th>Weighted In-Degree</th>
<th>Out-Degree</th>
<th>Weighted Out-Degree</th>
<th>Eigenvector</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>English</td>
<td>English</td>
<td>English</td>
<td>Indonesian</td>
<td>English</td>
</tr>
<tr>
<td>2</td>
<td>Indonesian</td>
<td>Russian</td>
<td>Spanish</td>
<td>Armenian</td>
<td>Indonesian</td>
</tr>
<tr>
<td>3</td>
<td>Finnish</td>
<td>Indonesian</td>
<td>Portuguese</td>
<td>Marathi</td>
<td>Spanish</td>
</tr>
<tr>
<td>4</td>
<td>Spanish</td>
<td>Persian</td>
<td>Indonesian</td>
<td>Oriya</td>
<td>Finnish</td>
</tr>
<tr>
<td>5</td>
<td>Estonian</td>
<td>Spanish</td>
<td>Italian</td>
<td>Gujarati</td>
<td>Portuguese</td>
</tr>
<tr>
<td>6</td>
<td>Portuguese</td>
<td>Hindi</td>
<td>French</td>
<td>Hindi</td>
<td>Italian</td>
</tr>
<tr>
<td>7</td>
<td>Tagalog</td>
<td>Tagalog</td>
<td>German</td>
<td>Tamil</td>
<td>French</td>
</tr>
<tr>
<td>8</td>
<td>Italian</td>
<td>Portuguese</td>
<td>Russian</td>
<td>Serbian</td>
<td>Tagalog</td>
</tr>
<tr>
<td>9</td>
<td>French</td>
<td>Polish</td>
<td>Turkish</td>
<td>Thai</td>
<td>German</td>
</tr>
<tr>
<td>10</td>
<td>German</td>
<td>Vietnamese</td>
<td>Hindi</td>
<td>Ukrainian</td>
<td>Estonian</td>
</tr>
</tbody>
</table>

language according to the CIA’s World Factbook [21]. For instance, the language distribution in Canada is 59% English, 22% French, etc., so English is its major language. Consequently, Canada is yellow in the map since English is in the yellow community in the LN. India is colored by states, since it has multiple major languages (defined at state level [55]), and they are located in different communities. The map shows neighboring areas (contiguous) with predominantly same colors, suggesting that the LN embeds some notion of geographical proximity.

4.2.3.3 Language Characterizations

The language network gives the overview of language interactions. In order to get a more specific understanding about a language, we can use a visual representation of it. Plots of the rows of matrix $K$ (language probability distributions) are very informative, especially when presented as ranked histograms.

A sample of language characterizations is shown in Figure 4.11 (full picture in Appendix B.5). For instance, among Dutch users, around 2%, 0.3%, and 0.04% also tweet in English, German, and French, respectively. English is the second
Figure 4.10: Revealing language relationships via network. (a) The language network based on the Olympics16 dataset – nodes are languages, colored based on their communities, and sized by the weighted in-degree; directed links colored as the source node representing the percentage of users in the source language that also use the target language. (b) The spatial dependence captured when grouping languages; countries are colored according to the community of their major language.
preferred language for the majority of users of all languages, except for Bulgarian, Pashto, and Ukrainian users. Clearly, there are distinct levels of multilingualism among languages; for example, Pashto users tend to, collectively, communicate in at most 3 languages, while English users collectively use more than 14. The “shape” of the language distributions can be used to explore these differences.

One can notice some similarities, even by highlighting only the 3-most used languages per plot. For instance, in Spanish and Portuguese, the $2^{nd}$ and $4^{th}$ languages are English and Italian, respectively, while they are both the $3^{rd}$ option for each other. Moreover, the distribution of the languages are not always symmetric, such as Spanish–Portuguese or Thai–Korean. German is more important to Dutch users than Dutch is for German users. Similarly, Ukrainian users prioritize more Russian than the other way round.

In section 4.2.3.1, we observed the locality effect by simple comparing the correlation of the number of speakers in languages and pairs of languages among datasets. We can use the city-level datasets to better visualize this effect. Figure
Figure 4.12: The locality effect captured by language distributions. Languages characterized in rows—English, French, and Japanese, respectively; and across city-level datasets. For instance, the second language of English speakers is Japanese, Spanish, or French if you live in Tokyo, NYC, or Paris, respectively.

4.12 shows how different English, French, and Japanese are when their speakers live in Tokyo, New York City, and Paris. Their secondary languages are not the same, but dependent of their neighborhood. The full language characterization of these datasets can be found in appendixes B.6, B.7, and B.8.

4.2.3.4 Language Clusters

The language distributions reveal more details about individual languages than the language network; they can more easily show similarities among languages. We investigated the extent of similarity among the languages by using the Agglomerative Clustering algorithm [121]. The hierarchical clusters can provide additional information between the specificity of a language distribution and the generality of a language network. The elements to be clustered are languages (rows of matrix
Figure 4.13: Revealing language relationships via clusters. The language hierarchical clusters for datasets at different locality-levels: (a) Tokyo, (b) The UK, and (c) Olympics. The similarity matrix used to create the hierarchical cluster (c) is also shown. (d) The spatial dependence captured when clustering languages using a global dataset; countries are colored according to the clusters of their major language in (c). Groups also tend to fill contiguous neighboring as in Figure 4.10.
where each component represents the probability that a user of a language communicates using another language. As proposed in Section 3.3.4, we used the Bhattacharyya distance as the affinity (distance) metric. We also tested the Cophenetic Correlation Coefficient and determined that “average” (the UPGMA algorithm) was the best linkage method [145].

Figure 4.13 depicts the hierarchical cluster for three datasets: Tokyo (4.13(a)), the United Kingdom (4.13(b)), and the Olympics16 (4.13(c)). Despite the differences between the datasets, some clusters remain consistent regardless the locality difference among datasets, such as Portuguese–Spanish; Chinese–Japanese; Norwegian–Danish–Swedish; and Dutch–German. However, to have a more general representation of language similarities, rather than local idiosyncrasies, we focus on the clustering of the Olympics16 dataset (Figure 4.13(c)).

Similarly as for the map colored based on the language network, Figure 4.13(d) shows a world map where countries are colored based on clusters of languages as defined in Figure 4.13(c). Upon close inspection, the map shows neighboring countries belonging to the same cluster, equivalent to the contiguous coloring pattern obtained from communities (Figure 4.10(b)); the spatial correlation is observed in both analysis, even when using very distinct methods to analyze the language characterizations (clusters and communities).

The visual effect of the spatial correlation is barely changed, regardless of the fact that clusters tend to be smaller than communities. For instance, although Telugu–Kannada and Tamil–Sinhala are two different clusters, all four languages are adjacent in the map and also belong to the same community on the LN. Yet, there are differences between the two approaches, for instance German–Dutch and Czech–Slovenian–Serbian share clusters, but in the community approach they are
not together, even though the countries where the language is mainly spoken share some geographical proximity.

The results showed languages are not grouped in communities or clusters based solely on syntactic similarity (family trees, see Figure 4.6). Groups are composed by languages widely spread in the language family tree structure, such as: the different-family group formed by Chinese (Sino-Tibetan), Japanese (Japonic), Korean (Koreanic), Thai (Tai-Kadai), and Vietnamese (Austroasiatic); or the same-family-different-branch group formed by Indo-European languages such as Persian and Pashto (Iranian branch) and Urdu (Indo-Aryan branch).

The syntactic diversity of languages found in groups by independent methods (communities and clustering), and the consistent spatial correlation patterns shown in the maps, suggest other factors beyond language family proximity drive people in their choices for interaction and social dependencies.

4.2.4 Discussion

In this work, we classified languages based on their relationship with other languages. Each language is represented as the emergent behavior from its preferred users. Users, on the other hand, are characterized based on the frequency of languages used in their tweets. We used 12 different datasets to understand whether the locality of a language had impact on its characterization.

Indeed, our results suggest relationships between languages are not limited to their origins (language family-tree), but they are strongly dependent on spatial factors (e.g., sharing borders). The support for this argument is two-fold. First, local datasets correlate in a significant lower degree to each other when compared to global datasets; if family branches were dominant they should correlate. Second,
the contiguous groups of adjacent countries are quite similar for both techniques used in the classification (clustering or community detection).

The characterization presented here has limitations since we cannot overcome possible bias embedded in Twitter data [99]. However, it is worthwhile noting that our characterization is centered on individuals, regardless of them being multilingual or their tweets being geo-tagged. The language characterization itself is a consequence of how we choose to aggregate users. Consequently, the results presented here tend to be less biased than those whose characterization considers multilingual users only [100, 130]. Yet, we plan to evaluate the precision of the Twitter language detection in our datasets by comparing the results against other tools, such as the chromium compact language detector.

We plan to do a characterization of places (based on their languages). This would require a simple redefinition of the membership function (Equation 3.5); we expect to have results on this approach in the very near future.

We believe our approach can be used to sense populations in real time, possibly detecting abrupt cultural shifts, such as in the presence of massive displacement of refugees. More importantly, the approach allows us to understand language barriers formed in social media. This can be useful when working with information spread applications (e.g., marketing campaigns, social network interventions). Our approach appears to indicate that information is likely to spread in accordance to spatial location of the countries more than based on language family branches.
4.3 Characterizing Organ Donation Awareness

Approximately 22 people die every day in the USA due to a lack of organs for transplant. Research suggests that the most effective solution is to increase organ donor rates; current, proposals range from expanding the donor eligibility criteria (donor pool) to performing mass media campaigns. However, little is known about the extent in which activities on social media are associated with aspects (e.g., awareness) of organ donation. Our hypothesis is that social media can be utilized as a sensor to characterize organ donation awareness and population engagement in donation for each different organ. In this sense, we collected tweets regarding organ donation, and characterized organ awareness by aggregating tweets from users who mostly mentioned that organ. Similarly, we assessed the relative risk (RR) between the cumulative incidence of organ-related conversations inside and outside geographical regions to characterize them regarding organ donation awareness. Our characterization suggests that organ-related conversations on social media seems to be indeed associated with aspects of organ donation such as the co-occurrence of organ transplantations. Also, we found variations regarding the specific organs that are prominently discussed in each geographical region, and that such variations seem to be associated with aspects of organ donation in that region; for instance, the abnormal amount of conversations about kidneys in Kansas. Our findings suggest that the proposed approach has the potential to characterize the awareness of organ donation in real-time. Our results were published in the 2017 IEEE 33rd International Conference on Data Engineering (ICDE) [113].
4.3.1 Motivation

Organ transplantation saves thousands of lives every year in the USA [111] and around the world. Despite starting as an experimental medical procedure when the first organ was transplanted in 1954 [96], organ transplantation has become a reliable, effective, and the preferred alternative for end-stage organ failure. Unfortunately, organ transplantation only reaches a small fraction of transplant candidates, and nearly 22 patients die in the USA every day for not having access to a transplant organ. As an example, in 2012, although roughly 60 thousand patients were in the waiting list for a kidney transplant in the USA, only 17 thousand kidney transplants were performed; less than $1/3$ of what was needed.

One of the approaches that have been proposed to improve the shortage of organs focuses on the expansion of the criteria for becoming an eligible donor [103] by increasing the use of higher risk grafts; however, this approach presents side effects such as delayed graft function of the transplanted organ with potential compromise of the graft function both short-and long-term [85]. To better inform policy makers, many past research efforts have focused on the assessment of the organ allocation process. Some works investigated the major factors associated with the survivability of transplanted organs such as the ischemic time [124], and found that the allocation needs to be tailored for each organ. Other works attempted to understand the complex network structure of organ transplantation using a geographic social networks [162], and found some geographical disproportion between donors and recipients as well as some anomalies regarding different organs in the allocation process. Ultimately, the change in allocation policies was debated aiming at reducing regional organs accessibility disparities [36].

Although the aforementioned research efforts attained significant results, they
also point to a research agenda focusing on raising the number of donors [17]. In this sense, conversation is a particularly important issue to organ donation awareness [15]. Establishing an effective conversation with families of donor-eligible patients may improve families’ consent rates [141] especially when families are aware of organ donation. Commonly, families are approached near the death of their loved ones, and approximately half of the families tend to refuse the request for donation [142]. Besides, families are also more likely to authorize donation if they had previously discussed organ donation with the deceased, and they knew the deceased’s wishes regarding organ donation [142].

Social media has evolved as a new tool to deal with the organ donation issue because it is cost-effective, and it has the potential to attain a higher population outreach [17]. In fact, there has been evidence that people look for social media as a way to create support groups, and that their conversations may lead to a structured social network [1]. Social media applications range from identifying potential kidney donors [22], helping organ donation advocacy agencies to increase online social network engagement [5], and increasing donor registration rates [17]. Yet, despite considerable research on organ donation using social media [5, 22, 28, 15], little has been done to associate activities in social media with real-world aspects (e.g., statistics) regarding organ donation.

This application aims at exploring the extent in which social media, such as Twitter, may be used to sense the population regarding organ donation awareness. This is the third and last instantiation of the methodology proposed on Chapter 3 in this dissertation. Our characterization demonstrates that social media has sufficient information regarding organ donation awareness, and has the potential to be employed as a social sensor for organ donation campaigns.
4.3.2 Experiment Settings

4.3.2.1 Methodology Instantiation

In this section, we detail how the proposed methodology was instantiated in the organ donation context.

- **Data** – a Twitter dataset.

- **Entity** – an organ.

- **Mention** – a co-occurrence of a context word (e.g., donation and transplantation) and an organ (see Figure 4.14) in the tweet text.

- **Set of Entities** – the six major solid transplanted organs in the US: heart, kidney, lung, liver, pancreas, and intestine.

- **Aggregation Methods** – preferred entity, with the group-characterization representing organ characterizations; and regions, with group-characterization representing the US states.

- **Analyses** – ranking, visualizing, highlighting, and labeled grouping (users and states).

4.3.2.2 Characterizing Users, Organs, and States

To build our social sensor for organ donation, we need to capture singularities not only among different organs, but also among distinct places. The first step is to define a structured approach to characterize the continuous and seemingly random stream of information from Twitter.
A straightforward approach is to build a characterization model based on single messages. Despite its intuitiveness, such characterization may be biased by the existence of a few heavily-active users. Twitter has a very heterogeneous tweeting rates per user, ranging from hundreds tweets per day to a handful in months. In addition, users are more likely to mention multiple organs within several tweets than a single tweet contain multiple mentions (see Figure 4.15(b)). Thus, to better represent the population, a characterization based on users is more appropriated. Ultimately, a user will be a representation of a collection of tweets.

As proposed on Section 3.2.1, users are defined based on the amount of attention they give to a set of entities—here, the set of most common transplanted solid organs. More specifically, we measure user’s attention from his/her tweets as frequencies of mention to organs in the donation context. Formally, we represent \( m \) users and their respective attention to \( n \) organs using a normalized contingency matrix \( \hat{U} = [\hat{u}_{ij}]_{m \times n} \). In this matrix representation, each row fully represents a user, i.e., \( \sum_{j=1}^{n} \hat{u}_{ij} = 1 \).

Individual characterizations may be too specific to perform an exploratory data analysis. Therefore, to help us to learn and to gain intuitions about the data, we aggregate users to explore two perspectives (see Section 3.2.2):

**Organs:** This hypothesis is that the characterization of an organ in relation to a set of organs can detect dependencies among them, such as the co-occurrence of transplantation (two organs are transplanted at the same time) and the cascade effects in organs failure. We characterize users by the extent of their attention to different organs. We assume an organ can be represented by averaging the behavior of the subset of users who dedicate most attention to it. This approach follows the *preferred entity* aggregation.
**Regions:** This hypothesis is that the characterization of regions in relation to a set of organs may reveal differences among regions regarding health issues, local policies, or levels of engagement in the donation cause. A region is represented by the aggregation of their inhabitants.

The two characterization are obtained by defining different membership-indicator matrices $P$ (see Section 3.2.2). In the characterization of organs, $m$ users are aggregated based on their most cited organ and in the characterization of regions, $m$ users are aggregated based on their locations. Using each membership-indicator matrix $P$, we can finally derive our aggregation matrices $K$. In this sense, the interpretation of $K$ depends on the definition of $P$. For instance, if $P$ aggregates based on the most cited organ, then $K = [k_{ij}]_{n \times n}$ and rows in $K$ contain the characterization of $n$ organs. However, if $P$ splits users based on regions; then, $K = [k_{ij}]_{r \times n}$ and rows represent the characterization of $r$ regions.

Either characterizing users, organs, or regions, we may ask:

1. To what degree can social media be used to study organ donation awareness?
2. Is there any difference in how US states relates with organ donation?
3. Are there preferences for specific organs?
4. Can we find groups of users based on mentioned organs?

**4.3.2.3 Dataset**

We collected data from Twitter because it is one of the most popular social media in the United States, and it is commonly used by researchers in social experiments. More importantly, Twitter allows data collection for virtually any of its users given
users tend to leave their profile public, i.e., everyone can read their posts. Our processing pipeline has three steps. First, tweets are collected using a filter based on predefined organ donation predicates (keywords). Then, the collected tweets are augmented to include their location; this can be done using the tweet geo-tag or the user location found in his profile. Finally, the augmented tweets are filtered again to retain only those belonging to USA users.

In order to focus on conversations regarding organ donation, we constrained our search with a set of keywords \( Q \) that are used to filter tweets using the Twitter Stream API. Figure 4.14 shows \( Q \) as the Cartesian product of a set of Context words (limited to organ donation terms) and a set of Subject words (limited to organs of interest). This approach guarantees every collected tweet in our dataset contains at least one word from Context and at least one of the words from Subject. Therefore, our dataset is conceived in the context of organ donation.

\[
Q = \left\{ \begin{array}{c}
\text{Context} \\
\text{transplant} \\
\text{transplantation} \\
\text{donor} \\
\text{donation} \\
\text{donate}
\end{array} \right\} \times \left\{ \begin{array}{c}
\text{Subject} \\
\text{heart} \\
\text{kidney} \\
\text{liver} \\
\text{lung} \\
\text{pancreas} \\
\text{intestine}
\end{array} \right\}
\]

Figure 4.14: The set of keywords used to collect tweets related to organ donation awareness is the Cartesian product of Context and Subject words.

This work focuses on characterizations within the USA, so we only kept the tweets from users located in the USA. Most common options to identify the geographical location of a Twitter user is to use the GPS coordinates (included in some tweets), or to use the self-reported location field in the user profile. GPS
Table 4.7: Statistics of the dataset used in the organ characterization.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start Data Collection</td>
<td>Apr 22\textsuperscript{th} 2015</td>
</tr>
<tr>
<td>Finish Data Collection</td>
<td>May 11\textsuperscript{th} 2016</td>
</tr>
<tr>
<td>Number of Days</td>
<td>385</td>
</tr>
<tr>
<td>Tweets collected</td>
<td>134,986</td>
</tr>
<tr>
<td>Number of Users</td>
<td>71,947</td>
</tr>
<tr>
<td>Avg. Tweets / Day</td>
<td>350</td>
</tr>
<tr>
<td>Avg. Tweets / User</td>
<td>1.88</td>
</tr>
<tr>
<td>Organs mentioned / Tweet</td>
<td>1.03</td>
</tr>
<tr>
<td>Organs mentioned / User</td>
<td>1.13</td>
</tr>
</tbody>
</table>

*134,986 out of 975,021 tweets could be identified as from USA users.

coordinates are more precise and dynamic, but much rarer (about 1.4% [102]). The user profile info is more static and abundant, but requires extra computation and is less precise. In this work, we locate users (country/state) augmenting their self-reported location using OpenStreetMap\textsuperscript{5}. This method has been shown to be reliable even at the county level [99].

Finally, our dataset comprises one year data and represents more than 70 thousands users in the USA; Table 4.7 summarizes the statistics. Figure 4.15(a) shows the number of users mentioning each organ. For instance, heart and intestine are, respectively, the most and least mentioned organs. The attention given to organs on Twitter highly correlates with the number of transplants in the USA (Spearman correlation, $r = .84$, $p < .05$); except for heart, first in popularity on Twitter, but third on number of transplants. Figure 4.15(b) shows the comparison between the number of tweets and the number of users mentioning multiple organs. The number of tweets is greater than the number of users only for single mentions.

\textsuperscript{5}http://www.openstreetmap.org

91
4.3.3 Results

In this section, we present the results of characterizing users based on organ transplantation conversations on Twitter from different perspectives of aggregation. First, we investigate the distribution of mentions to organs in the USA in order to grasp intuitions about them. Second, we analyze the states in the USA identifying organs highlighted in conversations, and revealing underlying similarities among them. Last, we explore the possibility of grouping users as a first step in the direction to identify common topics in the conversations of organ donation.

4.3.3.1 Organ Perspective

The organ characterization is based on the aggregation of users who behave similarly, i.e., users who mostly mention the same organ. Figure 4.16 shows the characterization of the six major organs transplanted in the United States; each plot represents an organ.

In order to emphasize the co-occurrence differences, the amount of attention is presented in ranked bins, from left to right. For instance, liver (Figure 4.16(b))
Figure 4.16: Characterization of the six major solid organs based on Twitter conversations. Each plot represents an organ, i.e., a row on $K$; it conveys how the user whose primary focus is on a specific organ also mentions other organs. The information for (a) heart, (b) liver, (c) pancreas, (d) intestine, (e) kidney, and (f) lung is depicted in red, green, olive, magenta, yellow, and blue, respectively. Note that the histogram bars are in log scale and their values are ranked.

Tend to be mentioned more frequent with kidney, heart, and lung, respectively. Kidney is the most important organ for heart, liver, and pancreas. Heart, on the other hand, is more important for intestine, kidney, and lung. Clearly, these co-occurrences are not reciprocal.

The organ characterization can show dependencies among them, such as dual organ transplantation. For instance, dual organ transplantation is more common among the pairs heart–kidney, liver–kidney, and pancreas–kidney [111]. Although further analysis are needed, the results in Figure 4.16 show how the population perceive the combined importance of organs. Another plausible explanation relates to the awareness of users who are interested in one organ transplantation with other types of transplantation. Such understanding can help us to have more effective
social network intervention strategies. For instance, users who are more aware of lung transplant may be more influenced to get involved in programs related to heart transplant than kidney transplant (Figure 4.16(f)).

Furthermore, the relations shown in Figure 4.16 might also indicate side effects, or how an organ failure can lead to other organs failures [66]. People who have heart disease can have renal dysfunction which is commonly caused by diabetes and hypertension [138]. Similarly, people with heart disease develop fluid retention which damages the liver [139]. Then, a small portion of these patients needs a liver transplant and as heart and liver are damaged, the kidneys are usually affected. People who have liver disease tend to have renal dysfunction, some of them due to diabetes, but many others due to the liver disease dysfunction. Note that we are not arguing that the conversations indicate the co-occurrence of failures of different organs, but rather that because of these co-occurrence people may have a tendency to talk about them together.

Finally, the analysis of intestine is less significant, since the majority of transplants happen in pediatric patients, and they are only related to a small fraction of the overall organ transplants [111]. This fact leads to less reliable statistics. However, all these may reflect the coexistence of diseases and problems, or the level of awareness of individual inflicted/affected by problems in one or many organs.

4.3.3.2 Region Perspective

In this section, we explore the geographic characterization of regions, in this context, the USA states. Similarly to the organ characterization, a state is represented as a distribution of attention to the set of six organs. Figure 4.17 shows the characterization of all states and territories of the USA as histograms. Despite strong
Figure 4.17: Characterization of US states based on attention given to organs by their Twitter users. The states in the USA have different distributions of mentions to organs which might indicate awareness of programs and even links between states (when they have similar characteristics). The bins in the histograms indicate the intensities of attention given to each organ. These histograms have different “shapes”. For instance, most states in the USA have their first and second-most-mentioned organ as heart and kidney, which may indicate the overall “ubiquity” of these transplants [111].

similarities, every state appears to have its own histogram shape (organ signature).

We explore two aspects of this characterization:
Highlighted Organs – since the prevalence of organs mentioned is not normally distributed, we cannot perceive the highlighted organs by comparing absolute values of mentions. For instance, from Figure 4.17 we tend to believe all states highlight heart.

Overall Awareness – states seem to share underlying similarities when dealing with organs. Not only the importance rank of organs varies among states, but also in the amount of attention they give to each of them. For instance, apparently states can be split by their second most mentioned organ—kidney, liver, or lung; or by the number of significant organs mentioned (3-6).

4.3.3.2.1 Identifying Highlighted Organs per State

We want to understand the impact/correlation of different states according to organ-related conversations. This might allow us to understand spatial disparities regarding organ-related conversations, identify clustering of well-defined borders of adjacent regions and geographic anomalies. For instance, is there any particular state in the USA unexpectedly associated with a specific organ-related conversation?

The simplest approach to answer this question is to count the number of users mentioning each organ and use a “winner-takes-all” strategy, i.e., the organ most cited is the one highlighted for that state. However, since some organs are much more prevalent than others, it is more likely to find a greater number of users mentioning that organ everywhere. Figure 4.17 shows heart as the prevalent organ in all the states in the USA.

To minimize this problem, instead of using the prevalent organ in a state, we calculate the relative risk (RR) as described on Section 3.3.2. In this sense, the RR
gives the excessive incidence of an organ in a state *relative* to the overall incidence in the rest of the USA. Figure 4.18 shows the highlighted organs in each state. Although most of them have at least one organ highlighted, for some states there are no significant excess for any organ and, thus, no organ is emphasized there, while other states have more than one highlighted organ. Appendix Figure B.9 shows the RR for all organs in all states.

The correlations between health-related traits geographically are commonly analyzed. For instance, the higher risk of hypertension observed in the so-called *stroke belt* in Southern USA which is associated with diet. Similarly, the increase amount

![Map of US States colored according to excessive conversations about specific organs. The excess (relative risk - RR) is explained by deviations to other states (see Eq. 3.13). A state is colored by organs with significant RR (confidence interval lower limit is greater than zero). Three inset examples – Louisiana, Massachusetts, and Rhode Island – show organs’ RR which is depicted in blue when they are significantly highlighted (Figure B.9 for more details).](image-url)
of liver disease in the Western United States due to fatty liver, probably associated with diet and genetic traits. Our results, for instance, show that Louisiana is associated with excess of kidney conversations while Massachusetts with both kidney and lung. Similarly, previous work analyzing geographic patterns of end-stage renal disease, kidney transplantation and deceased donors, found Kansas as the only state with a surplus of deceased kidney donors [18] in Midwestern USA. Interestingly, Kansas is also the only state in the Midwestern USA for which conversations of kidney is highly exceeding the national expectation.

### 4.3.3.2.2 Clustering States in the USA

In addition to identifying highlighted organs, one might be interested to investigate similarities between states considering all organs. For instance, states can be similar not only based on organs that exceed national expectation, but they can also be similar according to the organs that are unexpectedly less mentioned.

The organ distributions reveal more details about states, showing some states to be more similar between each other. We investigated the extent of similarity between the states by using the Agglomerative Clustering algorithm [121]. The hierarchical clusters can provide additional information beyond organs highlight. The elements to be clustered are states (rows of matrix $K$), where each component represents the probability of mentioning an organ in that state.

Figure 4.19 shows the similarity matrix of states as a heatmap for which the lower values are associated with higher similarity. Using the dendrogram, the hierarchical clusters can be analyzed at any location on the hierarchy. Such clusters present some degree of consistence with the aforementioned results regarding the organs highlighted at each state (see Figure 4.18). For instance, Delaware, Rhode
Figure 4.19: Hierarchical clustering of states based on their similarity with regards to the extent of incidence of specific organ-related conversations. States are outlining zones of organ-related conversation. For instance, the states belonging to the cluster depicted in red are mostly associated with liver conversations.

Nebraska, Puerto Rico, US Virgin Islands, Delaware, Rhode Island, Colorado, Connecticut, Alaska, Tennessee, Louisiana, Mississippi, Montana, Utah, Hawaii, Northern Mariana Islands, Wyoming, New Hampshire, North Dakota, Idaho, Kansas, Massachusetts, South Dakota, Vermont, Illinois, New Jersey, Kentucky, Nevada, West Virginia, Pennsylvania, Alabama, Wisconsin, Arkansas, Maryland, New Mexico, Michigan, Arizona, Indiana, District of Columbia, New York, Oregon, Georgia, Virginia, Maine, Minnesota, South Carolina, Iowa, Oklahoma, Texas, Florida, North Carolina, Ohio, Washington, California, Missouri. Indeed, from the leftmost state to the rightmost state in the similarity matrix, Nebraska to Missouri, respectively, the states are outlining zones of organ-related conversation in the following order: liver (from Delaware to North Dakota), lung (from Massachusetts to Wisconsin), kidney (from New York to Virginia), and heart (from Minnesota to California). Similarly, states without a highlighted organ tend to cluster, for instance, in the zone between New Mexico and Indiana.
4.3.3.3 User Perspective

So far, we explored the relations among organs and regions strictly according to the maximum attention and state borders, respectively. This first two characterizations can be seen as a validation phase where we could detect the richness and accuracy of the information hidden on tweets. Since, individual user characterization does encode valuable information, we can also learn from an aggregated characterization of them.

In this sense, as a preliminary investigation, we used K-Means to cluster users by their full behavior; not only based on the most-cited organ. After some empirical analysis comparing the inertia, the average cluster size, and the silhouette coefficient, we chose $k = 12$. Since we are characterizing six organs, $k$ must be at least six in order to allow at least one cluster for each organ. Indeed, since we have approximately 72 thousand users, even the smaller cluster which is associated with 0.3% of users would still be related to roughly 2 thousand users. Therefore, a greater number of clusters could still be used.

Figure 4.20 brings the characterization of each cluster, as well as their relative size. Although these clusters still demand more investigations, they already seem to reveal some interesting information allowing us to identify which users present the general patterns already identified for organs (see Figure 4.16). We can identify the subset of users focusing on a single organ, and also users focusing on two and three organs. These clusters might even represent organ-related users with different attitudes towards organ donation. For instance, the bottom-rightmost cluster of user (see Figure 4.20) mention virtually all organs especially when compared with the other clusters. This information can be used, again in conjunction with region, to investigate possible further correlations.
Figure 4.20: Cluster of users based on their conversations on Twitter using K-Means and their relative size. We chose $k = 12$ clusters based on the silhouette coefficient, average cluster size and inertia which were 0.953, 31697.42 and 2512.27, respectively. These clusters shows possible classification of users and might be related to different users’ roles in the organ donation environment.

4.3.4 Discussion

In this application, we characterized social media users and states in the USA based on their attention to different solid organs; we use markers (i.e., indicators) of awareness, norms, and behaviors towards organ donation. This characterization might lead to a better understanding of these users and their geographic variations. For instance, the geographic characterization of organ-related conversation at the state level can help us identify patterns of awareness from the angle the states in the USA. Similarly, our characterization might be used to differentiate classes of users such as health care practitioners, donors, waiting-list candidates, organ donation advocacy agencies, or simply demonstrate that different users have different behaviors towards organ donation.

The potential impact of this characterization is that it can improve the assessment of organ donation awareness approaches in the United States, but also
derive social intervention approaches that better fit the cultural, religious, and educational differences between states. Ultimately, this characterization can inform models of social influence to be employed in the context of organ donation aiming at designing interventions that effectively target specific groups of users.

Possible limitations of our work regards to bias in the collected data. The population of the United States is underrepresented by Twitter users since they are a highly non-uniform sample of the USA population especially with regards to geography, gender and race/ethnicity [99]. Twitter users are biased towards highly populated counties and male users. Also, depending on the region, different race/ethnicity (i.e Caucasian, African-American, Asia and Hispanic) can be over-sampled or under-sampled. For instance, the Midwestern population of United States is underrepresented among Twitter users.

We expect works such as ours to play an important role in social awareness programs because it enable researchers to design targeted approaches that fit the diversity in social networks. The power of social influence can be better harnessed when we have a better understanding of the complex structure that is formed when individuals interact with one another in online social networks.
Chapter 5

Learning from Characterizations

In this chapter, we show two different applications of how the \textit{a posteriori} knowledge generated in Chapter 4 can be used as \textit{a priori} knowledge in further research. Check Figure 5.1 to perceive this in the context of the methodology. Particularly, we used the diversity of football supporters in a region (Section 4.1) to estimate social disorganization [112], and the language network (Section 4.2) as a proxy for socio-economic development [134].

Figure 5.1: The knowledge acquired by analyzing the initial characterizations can be used in further investigations.
5.1 Estimating Social Disorganization

Urban communities can benefit from behavior regulation of their members in the interest of collective values. The absence of such control is related to the concept of social disorganization, and it is hypothesized to be associated with crime and anti-social behavior in neighborhoods. Social disorganization is, however, hard to quantify due to the lack of data and the inherent complexity that emerges from social interactions. Notably, geolocated social media provides a real-time assessment of places via the examination of the digital footprints left by users. In this dissertation, we introduce a measure for social disorganization by analyzing geotagged posts on Twitter. We propose to characterize the social disorganization of a place by evaluating the entropy of individuals’ opinions about certain subjects. As a case study, we used tweets related to football in the UK, given its ubiquity in that country, which makes its supporters as proxies for the social characteristics of those places. We found our measure can reasonably explain the variation of the occurrence of crime across regions in UK, and it better explains the variation of crime among places with higher social disorganization. Our results were published on the 30th International Florida Artificial Intelligence Research Society (FLAIRS) Conference [112].

5.1.1 Motivation

Crime takes place unevenly across places in cities. This characteristic of crime is argued by ecological criminologists to be the outcome of features in the social fabric of the places themselves [42]. Social ecologists highlight the active role of place in criminal occurrence, and consider crime as the product of social disorganization
which is produced by social changes, such as immigration, rural-urban migration, high social mobility, among others [135]. These changes undermine social arrangements, such as traditional control institutions, traditional stable structures, and established coping behavior [77]. A deficient social structure in a community drives people to compete rather than to cooperate. The residents of cooperative communities have the capability to organize themselves, and to share expectations for the social control of public space, a concept called collective efficacy [135]. The breakdown of such community control leads to a disorganized community insulated from conventional norms and prone to criminal activities [42].

The place-oriented view of crime advocates for an analysis of the social structure of communities and its relationship with offenses, but the lack of data has hindered a broader assessment of these theories. Notably, the increasing amount of localized data available from social media have the potential to help in these analyses. An example of this is Twitter, the platform has allowed researchers to understand and to model human behavior, tackling different aspects of crime [23, 50, 107, 117, 167, 164, 172].

In this dissertation, we propose a measure of social disorganization in places by examining geotagged tweets. We characterize a place using the entropy of different types of tweets in the region. In order to have a proxy of the social features in places, we gathered tweets about one of the most popular sports in the world: football (aka soccer). Due to its popularity, it is hard to imagine the behavior of the supporters as independent from other aspects of society, particularly in places such as Brazil, Italy, Germany, and the United Kingdom. We used football-related tweets from the UK to examine the relationship between our measure of social disorganization and criminal occurrence. For our measurement, we considered
mentions to clubs’ official Twitter accounts and hashtags in the tweets, similarly as described in Section 4.1.2. First, we evaluated the correlation between our proposed measure and the crime rate in places in the UK. Then, we isolated the population effect from users by performing a partial correlation. Finally, we built regression models to assess the contribution and the significance of our proposal to estimate social disorganization in the form of criminal activity.

5.1.2 Experiment Settings

5.1.2.1 Characterizing Regions

As proposed in 3.2.1 and implemented on 4.1.2, one can characterize users based on the amount of attention given by them to a set of football clubs; aggregate them based on their preferred entity; and, end up having the characterization of the clubs (entities). This work uses a distinct aggregation function to characterize a region as the emergent behavior of its users. A user belongs to all regions where he/she tweeted. Therefore, supporters (Twitter users) can be characterized in two ways:

**Globally** – one characterization per user based on all his/her tweets. Focus on a single user characterization, and how it influences the regions where he/she interacts.

**Locally** – a different characterization per region he/she has tweeted; focus on regions, potentially allowing users to behave differently over distinct places.

In this work, we use local characterization to understand the role of regions over users’ behavior and their social disorganization. Moreover, a region is a Lower
Layer Super Output Area (LSOA), which is a hierarchy proposed to improve the reporting of small area statistics about geographical areas of England and Wales. Thus, each type of event (crime occurrences or tweets) is aggregated by LSOA.

5.1.2.2 Datasets

Football

To characterize regions based on football supporters, we used a Twitter dataset capturing part of the 2014/2015 English Premier League. This is a “subset” of the dataset presented on Section 4.1.2.3 with two main differences: (i) we considered mention to clubs not only their verified Twitter accounts (@), but also their official hashtags (#); and (ii) we only considered geo-tagged tweets within the UK, more specifically, within the LSOA regions. Table 5.1 presents some statistics. Around 90,000 tweets were collected over a 3-month period, representing around 25,000 different users.

As for the previous applications, we do not need to apply natural language processing to identify mentions to clubs since accounts and hashtags are indexed by Twitter as specific metadata. Therefore, a mention is a simple lookup comparing these entities and the official terms for each club. Table 5.2 shows all terms used and which clubs were tracked.

Crime

The dataset containing the crime-related events that occurred in the UK was retrieved from the Open Data project (data.gov.uk) filtered by the period of the Twitter dataset, i.e., from February to May of 2015. The leftmost part of Table 5.3 describes the different types of crime in this dataset and the
Table 5.1: Statistics from the football Twitter dataset.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start Data Collection</td>
<td>07/Feb/15</td>
</tr>
<tr>
<td>Finish Data Collection</td>
<td>07/May/15</td>
</tr>
<tr>
<td>Number of Days</td>
<td>89</td>
</tr>
<tr>
<td>Tweets with Mentions</td>
<td>89,416</td>
</tr>
<tr>
<td>Users with Mentions</td>
<td>24,974</td>
</tr>
<tr>
<td>Tweets per User</td>
<td>3.58</td>
</tr>
<tr>
<td>Tweets per Day</td>
<td>1,005</td>
</tr>
</tbody>
</table>

numbers of crimes that occurred in the period considered.

5.1.2.3 The Entropy of Football Supporters

Our hypothesis is based on the social disorganization theory which states that crime is related to the disorganization of the society in a region [42]. Thus, we need to define social disorder in a football context. The rate of similarity among supporters may be a proxy for disorder. For instance, places with similar individuals are less prone to conflict than those with people with conflicting ideas. Hence, we argue that regions with single club supporters are more socially organized than places containing supporters from several clubs.

Thus, we use entropy to measure the level of social-football disorganization in a region. The supporter probability distribution of a region \( k_i \) (rows of \( K \), see Eq. 3.5) is the aggregation of all supporters’ behavior \( \hat{u}_i \) in that region. Therefore, we can calculate the normalized entropy \( \hat{S}_i \) of a region \( i \) as follows:

\[
S_i = -\sum_{j=1}^{n} k_{ij} \cdot \log k_{ij}
\]
Table 5.2: Official Twitter accounts and hashtags of the football clubs considered in our analysis.

<table>
<thead>
<tr>
<th>Club</th>
<th>Account</th>
<th>Hashtag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arsenal</td>
<td>@arsenal</td>
<td>#AFC</td>
</tr>
<tr>
<td>Aston Villa</td>
<td>@AVFCOfficial</td>
<td>#AVFC</td>
</tr>
<tr>
<td>Burnley</td>
<td>@BurnleyOfficial</td>
<td>#BURNLEYFC</td>
</tr>
<tr>
<td>Chelsea</td>
<td>@ChelseaFC</td>
<td>#CFC</td>
</tr>
<tr>
<td>Crystal Palace</td>
<td>@CPFC</td>
<td>#CPFC</td>
</tr>
<tr>
<td>Everton</td>
<td>@Everton</td>
<td>#EFC</td>
</tr>
<tr>
<td>Hull</td>
<td>@HullCity</td>
<td>#HCAFC</td>
</tr>
<tr>
<td>Leicester</td>
<td>@LCFC</td>
<td>#LCFC</td>
</tr>
<tr>
<td>Liverpool</td>
<td>@LFC</td>
<td>#LFC</td>
</tr>
<tr>
<td>Man City</td>
<td>@MCFC</td>
<td>#MCFC</td>
</tr>
<tr>
<td>Man Utd</td>
<td>@ManUtd</td>
<td>#MUFC</td>
</tr>
<tr>
<td>Newcastle</td>
<td>@NUFC</td>
<td>#NUFC</td>
</tr>
<tr>
<td>Qpr</td>
<td>@QPRFC</td>
<td>#QPR</td>
</tr>
<tr>
<td>Southampton</td>
<td>@SouthamptonFC</td>
<td>#SAINTSFC</td>
</tr>
<tr>
<td>Spurs</td>
<td>@SpursOfficial</td>
<td>#THFC</td>
</tr>
<tr>
<td>Stoke</td>
<td>@stokecity</td>
<td>#SCFC</td>
</tr>
<tr>
<td>Sunderland</td>
<td>@SunderlandAFC</td>
<td>#SAFC</td>
</tr>
<tr>
<td>Swansea</td>
<td>@SwansOfficial</td>
<td>#SCAFC</td>
</tr>
<tr>
<td>West Brom</td>
<td>@WBAFCofficial</td>
<td>#WBA</td>
</tr>
<tr>
<td>West Ham</td>
<td>@whufc_official</td>
<td>#WHUFC</td>
</tr>
</tbody>
</table>

\[
\hat{S}_i = \frac{S_i}{S^+}, \tag{5.1}
\]

where \(S^+\) is the maximum entropy that a region can assume. \(S^+\) happens when supporters are evenly distributed among all clubs (maximum social-football disorder), i.e., when \(k_{ij} = \frac{1}{n}\) for all \(j\) clubs.

### 5.1.3 Results

In this work, we proposed to measure the entropy of football supporters’ diversity in a region (Equation 5.1), and to analyze to which extent it can be used to measure social disorganization. Since social disorganization has already been shown to correlate with crime activity [42], we use football as a proxy for social...
disorganization and correlate it with crime.

Figure 5.2(a) depicts the spatial distribution of crimes in the UK during the period of our dataset by aggregating criminal events within each LSOA. The resident population size in each LSOA is normally distributed with an average (standard deviation) equal to 1584 (279). Still, the heatmap in this figure shows the existence of spots of high criminal rate, an aspect of crime supported by the literature of crime concentration [42]. Figure 5.2(b) shows the spatial distribution of the clubs with the most supporters in each region when tweets are aggregated in the same LSOAs. Although many clubs are the most popular in many different regions, some patterns can be seen by examining the clubs and their spatial neighbors, for instance: Newcastle (green) and Sunderland (purple) are prevalent in their regions in the Northern UK; Liverpool (red) and Everton (orange) are predominant in the Liverpool area; and there are some clubs quite popular everywhere on the map, like Arsenal (magenta) and Manchester United (yellow). Such simple analyses suggests

Figure 5.2: Plots of UK with inner plot from London area of: (a) crime, (b) regions labeled according to highlighted football clubs, and (c) entropy of football supporters.
the region characterization based on tweets is able to capture football supporters’ preferences. Moreover, Figure 5.2(c) shows the spatial distribution of the entropy of football supporters based on tweets in each LSOA with at least two users. Although the patterns in the entire map are not clear, the inner map (London area) presents visual similarities with the inset of the criminal map in Figure 5.2(a).

Due to the evidence that population size and crime rate are related [8], an estimator for the actual population size (i.e., transient population) in each place would probably already reveal a relationship with crime occurrence in a given region. The population information in each LSOA from censuses can not capture these dynamics, since census takes into account only the residents of the regions. In fact, Twitter has been used to find such transient population, an additional piece of information that improves crime prediction [89].

In order to address the contribution of supporters’ diversity on crime estimation, we need first to analyze how the number of tweets and users in a region correlate with the number of crimes. This would be the base for a null model that provides alternative information regarding tweets; this null model possibly is able to capture the influence of a transient population of regions on crime. To this extent, we calculated the correlations $\hat{\rho}_{TC}$, $\hat{\rho}_{UC}$, and $\hat{\rho}_{EC}$ between the number of tweets, the number of users, and the supporters’ entropy, respectively, with crime activity in each region, as shown in Table 5.3. The correlation values for entropy $\hat{\rho}_{EC}$ are higher than for users $\hat{\rho}_{UC}$, that are also higher than for tweets $\hat{\rho}_{TC}$, regardless of type of crime taken into account. Thus, apparently the number of users in a region explains more crime activity than the number of their tweets, and the entropy of supporter diversity explains crime even more than the number of users. However, this direct comparison may lead to wrong conclusions, especially
when the random variables could be correlated between each other. Thus, we also calculated the partial correlation $\hat{\rho}_{EC \cdot U}$ between the entropy of supporter diversity and crime activity, controlling for the effect of the number of users. The results show that the entropy correlation is not driven by a population effect.

The use of entropy as a proxy of disorder of a place and the correlation values found provide support to the social disorganization theory. However, due to the many aspects of crime and criminology, two points need to be raised: 

(i) social disorganization is not the only factor that explains crime, i.e., other factors can also drive the increase of crime in places with high or low levels of disorder; and

(ii) the places without disorder are also subject to criminal occurrences. Due to these intrinsic difficulties in the theories from criminology, we need to analyze user and entropy ranges in which crime rate is better explained.

The correlations shown in Table 5.3 considered regions with at least 5 users. Figure 5.3 addresses the impact of the minimum number of users in a region and the partial correlation $\hat{\rho}_{EC \cdot U}$ for all crimes. Intuitively, as we increase the minimum

![Figure 5.3: The partial correlation analysis $\hat{\rho}_{EC \cdot U}$ between all crimes and supporters entropy, controlling for number of users, increases as the minimum users increases, but the cover percentage of regions decreases.](image-url)
users’ constraint, the number of regions considered in the calculation decreases significantly. On the other hand, $\hat{\rho}_{ECU}$ increases as we keep the more populated regions, suggesting our proxy for social disorganization explains crime rates better in more populated areas, or at least, in regions more represented on Twitter.

Figure 5.4 depicts the relationship between entropy versus the amount of crime in each LSOA. This plot shows that entropy explains crime better when places with higher entropy are taken into account, a finding that holds true regardless of crime type. Yet, Figure 5.5 shows that the compound effect of increasing both users and entropy also reflects in increasing the partial correlation between all crimes and entropy of supporters. Therefore, our proxy works better explaining crime in places more populated and more disorganized.

In order to assess the explanation power of the entropy of supporters over crime, we constructed two linear models:

![Partial correlation $\rho_{CEU}$ with $U \geq 0$ and $E \geq e$](image)

Figure 5.4: The entropy of regions in the UK is positively correlated with crime rate, and this correlation is higher in more disorganized places, regardless of crime type. The white parts represent correlation with high p-values or the lack of points with a certain entropy value.
\[ C = \alpha + \beta_U U + \epsilon \] in which the number offenses in a region is explained by the number of Twitter users in that region.

\[ C = \alpha + \beta_U U + \beta_E E + \epsilon \] in which the number of offenses in a region is explained by the number of users and our measure of social disorganization.

Table 5.3 shows the adjusted R\(^2\) for both models and the regression coefficients of the second model. As expected, the contribution of number of users (\(\beta_U\)) is significant for most types of crime. Moreover, the contribution of the entropy of supporters (\(\beta_E\)) is significant regardless the type of crime, and its addition to regression causes a significant increase in the variance explained by the model (adjusted R\(^2\)). Nonetheless, we do not expect that football can explain all types and occurrences of crime, or that football is the sole component that leads to crime, since it is known that many other factors may lead to crime, such as population

![Figure 5.5](image-url)

Figure 5.5: Our proxy to social disorganization better explains crime rate in disorganized and populated places. The entropy of regions in UK is positively correlated with crime rate, and this correlation tends to increase as we filter locations by the amount of users and entropy. The white parts represent correlation with high p-values or the lack of points with a certain entropy value.
education level, social unrest, and economic opportunities, to name just a few [90].

Table 5.3: Statistics, correlations and regression summary per type of crime. The number of crimes within the period Feb/15 – May/15. Correlations and regressions considering regions with at least 5 users. The correlations $\hat{\rho}_{TC}$, $\hat{\rho}_{UC}$, and $\hat{\rho}_{EC}$ between the tweets count, the number of users, and the entropy of supporters in regions of UK, respectively, with criminal occurrence are increasingly stronger. The partial correlation $\hat{\rho}_{EC \cdot U}$ between entropy and crime whilst controlling the effect of the number of users in a region is always greater than or equal to $\hat{\rho}_{EC}$. The adjusted $R^2$ for the model $crime \sim user + entropy$ is significantly greater than for the model $crime \sim user$. Regression coefficients are the intercept $\alpha$, user $\beta_u$ and entropy $\beta_e$.

<table>
<thead>
<tr>
<th>Type of Crime</th>
<th># Crimes</th>
<th>$\hat{\rho}_{TC}$</th>
<th>$\hat{\rho}_{UC}$</th>
<th>$\hat{\rho}_{EC}$</th>
<th>$\hat{\rho}_{EC \cdot U}$</th>
<th>Adj. $R^2_{U \cdot E}$</th>
<th>Adj. $R^2_{U \cdot E + E}$</th>
<th>$\alpha$</th>
<th>$\beta_u$</th>
<th>$\beta_E$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anti-Social Behavior</td>
<td>593,238</td>
<td>0.06</td>
<td>0.12</td>
<td>0.24</td>
<td>0.25$^a$</td>
<td>0.013$^b$</td>
<td>0.073$^b$</td>
<td>-4.19</td>
<td>0.46$^b$</td>
<td>133.49$^b$</td>
</tr>
<tr>
<td>Bicycle Theft</td>
<td>26,177</td>
<td>0.01</td>
<td>0.06</td>
<td>0.23</td>
<td>0.25$^a$</td>
<td>0.002$^b$</td>
<td>0.063$^b$</td>
<td>-3.20</td>
<td>0.03$^a$</td>
<td>19.30$^a$</td>
</tr>
<tr>
<td>Burglary</td>
<td>129,974</td>
<td>0.01</td>
<td>0.07</td>
<td>0.26</td>
<td>0.26$^a$</td>
<td>0.003$^b$</td>
<td>0.070$^b$</td>
<td>2.53$^a$</td>
<td>0.03$^a$</td>
<td>13.42$^a$</td>
</tr>
<tr>
<td>Criminal Damage and Arson</td>
<td>170,953</td>
<td>0.05</td>
<td>0.10</td>
<td>0.24</td>
<td>0.24$^a$</td>
<td>0.008$^b$</td>
<td>0.065$^b$</td>
<td>3.52$^a$</td>
<td>0.05$^a$</td>
<td>17.82$^a$</td>
</tr>
<tr>
<td>Drugs</td>
<td>46,907</td>
<td><strong>0.08</strong></td>
<td><strong>0.15</strong></td>
<td>0.26</td>
<td>0.27$^b$</td>
<td><strong>0.023</strong>$^a$</td>
<td><strong>0.090</strong>$^b$</td>
<td>-2.20</td>
<td>0.08$^a$</td>
<td>18.68$^a$</td>
</tr>
<tr>
<td>Other Crime</td>
<td>18,582</td>
<td>0.04</td>
<td>0.09</td>
<td>0.19</td>
<td>0.19$^b$</td>
<td>0.007$^b$</td>
<td>0.043$^b$</td>
<td>-2.07$^a$</td>
<td>0.03$^a$</td>
<td>9.01$^a$</td>
</tr>
<tr>
<td>Other Theft</td>
<td>158,332</td>
<td>0.06</td>
<td>0.12</td>
<td><strong>0.38</strong></td>
<td><strong>0.39</strong>$^b$</td>
<td><strong>0.014</strong>$^b$</td>
<td><strong>0.163</strong>$^b$</td>
<td>-29.75</td>
<td>0.29$^a$</td>
<td>127.02$^a$</td>
</tr>
<tr>
<td>Possession of Weapons</td>
<td>7,214</td>
<td>0.03</td>
<td>0.06</td>
<td>0.17</td>
<td>0.17$^b$</td>
<td>0.002$^b$</td>
<td>0.031$^b$</td>
<td>-0.81$^b$</td>
<td>0.01$^a$</td>
<td>4.28$^a$</td>
</tr>
<tr>
<td>Public Order</td>
<td>57,969</td>
<td>0.06</td>
<td><strong>0.15</strong></td>
<td>0.31</td>
<td>0.32$^a$</td>
<td>0.021$^b$</td>
<td>0.121$^b$</td>
<td>-4.22$^a$</td>
<td>0.09$^a$</td>
<td>28.12$^a$</td>
</tr>
<tr>
<td>Robbery</td>
<td>16,245</td>
<td>0.04</td>
<td>0.09</td>
<td>0.25</td>
<td>0.26$^a$</td>
<td>0.008$^b$</td>
<td>0.073$^b$</td>
<td>-1.24$^a$</td>
<td>0.02$^a$</td>
<td>8.31$^a$</td>
</tr>
<tr>
<td>Shoplifting</td>
<td>111,714</td>
<td>0.05</td>
<td>0.12</td>
<td>0.26</td>
<td>0.26$^a$</td>
<td>0.013$^b$</td>
<td>0.080$^b$</td>
<td>-17.83$^a$</td>
<td>0.30$^a$</td>
<td>93.04$^a$</td>
</tr>
<tr>
<td>Theft From the Person</td>
<td>25,960</td>
<td><strong>0.08</strong></td>
<td>0.12</td>
<td>0.33</td>
<td>0.33$^a$</td>
<td><strong>0.014</strong>$^a$</td>
<td><strong>0.124</strong>$^a$</td>
<td>-19.83$^a$</td>
<td>0.18$^a$</td>
<td>68.84$^a$</td>
</tr>
<tr>
<td>Vehicle Crime</td>
<td>115,826</td>
<td>0.02</td>
<td>0.05</td>
<td>0.21</td>
<td>0.21$^a$</td>
<td>0.001$^b$</td>
<td>0.045$^b$</td>
<td>3.06$^a$</td>
<td>0.02$^a$</td>
<td>12.43$^a$</td>
</tr>
<tr>
<td>Violence and Sexual Offenses</td>
<td>304,348</td>
<td>0.05</td>
<td>0.12</td>
<td>0.32</td>
<td>0.33$^a$</td>
<td>0.014$^b$</td>
<td>0.122$^b$</td>
<td>-10.34$^a$</td>
<td>0.24$^a$</td>
<td>88.49$^a$</td>
</tr>
<tr>
<td>All Crimes</td>
<td>1,783,439</td>
<td>0.06</td>
<td>0.14</td>
<td>0.34</td>
<td>0.35$^b$</td>
<td>0.019$^b$</td>
<td>0.140$^b$</td>
<td>-80.62$^a$</td>
<td>1.85$^b$</td>
<td>642.26$^b$</td>
</tr>
</tbody>
</table>

Significance codes: $^a \rho < 0.10$, $^b \rho < 0.05$, and $^c \rho < 0.01$. For adjusted $R^2$, significance based on $F$ test.

5.1.4 Discussion

We proposed a measure for social disorganization in a region by analyzing the entropy of online social media data from users in that region. We carried out experiments using football supporters conversations and the diversity of clubs they mention in their geolocated tweets; more specifically, we used football-related tweets from the UK. We observed a significant correlation between the number of users and crime, and between the entropy of supporters and crime. Then, we measured
the partial correlation between them to confirm that the entropy correlation was not inflated by the population effect. Finally, we used regression models to confirm the contribution of both, the number of users and their entropy, to model crime activity. The coefficients for entropy are statistically significant, regardless the type of crime. For instance, when considering regions with at least 5 users and for all crimes, the model incorporating the entropy explains 7 times more the variance of crime (adj. $R^2$) than the model without it. We also found that our proposed measure of social disorganization explains better the variation of crime among regions with higher disorganization and larger population.

This work is a first attempt to create a framework to assess the levels of social disorganization in locations by using social media. It is worth noticing that this was the contribution here; the example with the UK football is a case study to demonstrate that certain subjects (in this case football) can be used to quantify social disorganization.

Although we found a positive correlation between our measure of social disorganization and crime, we aggregated longitudinal data in such way that the entropy of places and the amount of crimes were analyzed without taking into account their variations over time. This was the case mainly due to the temporal granularity of the data sets, i.e., monthly, provided by the police forces in the UK. Still, one of the benefits of our proposal is the capability to assess places in a real-time fashion; thus, as future work, and by the possession of richer criminal data sets, we want to find the minimum time window to extract significant characterizations of places, as well as to capture the movement of social disorganization in places over time and its relationship with crime mobility.

We also want to test the impact of characterizing regions based on a global
approach; that is, based on people’s behavior everywhere, and not on the behavior of people in a particular region (local approach as defined in Section 5.1.2.1). Moreover, we intend to examine other factors that can also be used to quantify social disorganization, and how they can aggregate value to the entropy of football in a predictive model. For instance, if we look at what people eat and calculate the entropy of their choices, would we get similar correlation numbers, or is there something special about football? How about the diversity of the languages spoken in a region? Can these seemingly unrelated social factors be combined in a calculation of social disorganization? It is generally difficult to get several datasets related to different factors for the same geographical regions. However, if such datasets are made available, the approach proposed here can be easily applied.
5.2 Indicating Socio-Economic Development

Globalization is a process driven by international trade leading to interactions and integration of people and government worldwide. Such process has been impacting us in many ways, including our living standards or quality of life (QoL). At the same time, the integration between people of the world leads to a stronger diffusion of languages, ideas, and values; more recently, this integration has received a further boost with the emergence of online social networks. Online social networks give us a platform to connect without any restriction of geographic regions, language, costumes, etc. In fact, they are a typical example of globalized world. Yet, the links between language usage in society and QoL has received little attention. Since both language usage and the QoL are influenced by globalization it is just natural that one tries to study both subjects combined. This work investigates if one can be used as a proxy for the other. Using approaches based on network science, our analysis of a large-scale Twitter dataset reveals that the patterns of user connectivities on online social networks (such as Twitter) as a function of languages usage is correlated to the QoL. Our results were published on the 30th International Florida Artificial Intelligence Research Society (FLAIRS) Conference [134].

5.2.1 Motivation

Globalization has been affecting many domains such as culture, economy, and international policies. The expansion of international trade and foreign investment is often seen as a sign of economic growth; higher economic growth rates and greater affluence are conducive to well-being. However, what is the well-being of a society? Many critics argue that gross domestic product (GDP) cannot capture the nation’s
of well-being even when based on purchasing power parity (PPP) [97]. They claim that GDP is intended to measure the productivity of a nation, and hence, it is an insufficient measure to quantify its quality of life (QoL), i.e., the well-being of a nation is not the economic growth alone, but a multi-dimensional concept. Which other social conditions play a role in improving QoL? Can the linguistic imperialism caused by the cultural globalization impact societies development?

In this last work, we look at how these two concepts (language and QoL) are intertwined; we are interested in understanding the extent to which language can be seen as a proxy of well-being, or if the distribution of languages used by people correlate with their socio-economic development.

The United Nations (UN) created the Human Development Index (HDI) as an alternative measure of prosperity embracing more than simple economic growth. The HDI is a composite metric that considers life expectancy, education levels, and per-capita income in its equation; thus, it is a better indicator of QoL of people than GDP. Although GDP is sometimes used as an indicator of prosperity, it does not always correlate with HDI. For instance, Cuba has low GDP per capita (PPP) US$6,876, but very high HDI of 0.863 (the max is 1) [136].

Authors have tried to look at relations between QoL and other factors. Ranis found that human development has important effects on economic growth [127]. An increase in the capabilities available to individuals allow them to pursue occupations in which they are more productive. In this sense, human development is correlated to the human capital and human capital, in turn, correlates with the economic growth. Several studies used HDI in an attempt to understand specific human characteristics. One of the more interesting examples look at the relation between HDI and obesity levels [95].
Mocanu et al. characterized the worldwide linguistic diversity in Twitter using geo-tagged data at different scales from country to neighborhood scales [100]. They show that the usage of Twitter is not uniform and has a correlation with economic factors. In another study, Kulshrestha et al. demonstrated the influence of geography in the cultural and linguistic backgrounds in Twitter [78]. Economic imbalances in society is correlated to the imbalance in the total number of tweets. For example, US accounts for 25% world GDP and 72% of all tweets.

Ronen et al. proposed to create global language networks (GLNs) using data from book translations, Wikipedia, and Twitter [130]. First, they counted the number of users interacting in pairs of languages; second, they identified which pairs appeared more than would be expected at random; then, they connected the paired languages forming the GLN. They suggested the fame of people to be related with the centrality of their native language in the GLN. Saha and Menezes demonstrated that the positions of languages on Twitter indicate interesting insights, including the visibility of information generated in a particular language [132, 133].

Although the observation of the connection between language and economic growth is interesting, more intriguing would be to see if languages used by people on Twitter can act as a proxy to their development scale. Twitter data has gained a lot of interest among the research scientists who are trying to understand the specific aspects of human behavior because it can act as a large real-time sensor of society [32]. Our main contribution in this work is to show the relation between the position of the languages in networks generated from datasets extracted from Twitter and the human development of the language (calculated as the function of the countries using that language). We validated our findings using statistical methods.
5.2.2 Experimenter Settings

5.2.2.1 Language Networks

As mentioned in the beginning of this chapter, the works presented here use the knowledge generated by the applications presented on Chapter 4. In this work, we used language networks generated as described on Section 4.2 based on global datasets – Olympics16 and G20 (for more details about the datasets, refer Table 4.5 and Appendix C).

5.2.2.2 HDI of Languages

The Human Development Index is a composite measurement for countries, but we are trying to understand the relationship between languages and HDI. Thus, before proceed to any network analysis we need to estimate the HDI of a language based on the language distribution of countries.

We propose the HDI of a language to be the average contribution of a single speaker of the language towards the world HDI, therefore summing the contributions of all the speakers of the language to the HDI of every country and then dividing by the total number of speakers. We collected the Human Development Index of countries as reported by United Nations [160], their language distributions in [130], the percentage of speakers per country per language, and countries’ total population as reported in the World Factbook by the Central Intelligence Agency [21]. Then, we calculated the HDI of a language $\ell$ as the weighted sum

$$\text{HDI}_\ell = \sum_c \text{HDI}_c \times P(c|\ell), \quad (5.2)$$
\[ N^\ell = \sum_c N^\ell_c, \quad (5.3) \]

\[ P(c|\ell) = \frac{N^\ell_c}{N^\ell}, \quad (5.4) \]

\[ \sigma(\text{HDI}_\ell) = \sqrt{\sum_c (\text{HDI}_c - \text{HDI}_\ell)^2 \times P(c|\ell)}, \quad (5.5) \]

where HDI_c is the HDI of country c, \( N^\ell \) is the number of speakers of \( \ell \) worldwide, \( N^\ell_c \) is the number of speakers of \( \ell \) in country \( c \), and \( P(c|\ell) \) is the probability of having a speaker from \( c \) speaking \( \ell \). The HDI of the languages are approximations because the values depend on many different factors, for instance, we are assuming HDI_c is invariant for the whole population of a country. The standard deviation of HDI_\ell is given by Equation 5.5.

### 5.2.3 Results

#### 5.2.3.1 Languages by HDI

In order to analyze the relatedness of the languages and the development of the countries where they are spoken, we used Equation 5.2 to calculate the HDI_\ell of all the languages that are present in our datasets.

Table 5.4 depicts a sample of countries and languages with highest and lowest HDI_c and HDI_\ell, respectively. For instance, Norwegian language has the highest HDI_\ell while those tweeting in Haitian has the lowest. Norwegian is dominantly spoken in Norway (which tops the HDI ranks for 12 years [129]), and Haitian is
Table 5.4: Highest and lowest ranked countries and languages according to their HDI; HDI<sub>c</sub> extracted from UN [160] and HDI<sub>ℓ</sub>, calculated using Equation 5.2.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Country</th>
<th>HDI&lt;sub&gt;c&lt;/sub&gt;</th>
<th>Language</th>
<th>HDI&lt;sub&gt;ℓ&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Norway</td>
<td>0.94</td>
<td>Norwegian</td>
<td>0.94</td>
</tr>
<tr>
<td>2</td>
<td>Australia</td>
<td>0.94</td>
<td>German</td>
<td>0.91</td>
</tr>
<tr>
<td>3</td>
<td>Switzerland</td>
<td>0.93</td>
<td>Dutch</td>
<td>0.91</td>
</tr>
<tr>
<td>4</td>
<td>Denmark</td>
<td>0.92</td>
<td>Swedish</td>
<td>0.91</td>
</tr>
<tr>
<td>5</td>
<td>Netherlands</td>
<td>0.92</td>
<td>Danish</td>
<td>0.91</td>
</tr>
<tr>
<td>-5</td>
<td>Burundi</td>
<td>0.40</td>
<td>Urdu</td>
<td>0.59</td>
</tr>
<tr>
<td>-4</td>
<td>Chad</td>
<td>0.39</td>
<td>Khmer</td>
<td>0.56</td>
</tr>
<tr>
<td>-3</td>
<td>Eritrea</td>
<td>0.39</td>
<td>Nepali</td>
<td>0.55</td>
</tr>
<tr>
<td>-2</td>
<td>Central African Republic</td>
<td>0.35</td>
<td>Punjabi</td>
<td>0.55</td>
</tr>
<tr>
<td>-1</td>
<td>Niger</td>
<td>0.35</td>
<td>Haitian</td>
<td>0.48</td>
</tr>
</tbody>
</table>

dominantly spoken in Haiti. Hence, a user tweeting in Norwegian is likely to have a better standard of living than a user tweeting in Haitian.

5.2.3.2 Correlation with Network Properties

The hypothesis we want to test is whether the language of an individual is correlated to his/her quality of life, i.e., if languages can be seen as catalysts of development by increasing or decreasing the access to information. Then, we used the language networks generated using Twitter data from two global datasets (G20 and Olympics) on Section 4.2 to calculate languages importance. In the network context, centralities can be seen as importance measures. We analyzed the most known centralities: in-degree, out-degree, betweenness, closeness, eigenvector, weighted in-degree, and weighted out-degree; each of them has its own specificity and can measure different perceptions of importance, such as popularity, power, proximity, etc. Table 5.5 lists the top 5 languages according to the
Table 5.5: Top languages ranked in the G20 and Olympics datasets according to in-degree and eigenvector centralities measured from the respectively language networks.

<table>
<thead>
<tr>
<th>G20 Dataset</th>
<th>Olympics Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-degree</td>
<td>Eigenvector</td>
</tr>
<tr>
<td>English</td>
<td>English</td>
</tr>
<tr>
<td>Indonesian</td>
<td>Indonesian</td>
</tr>
<tr>
<td>Bosnian</td>
<td>Spanish</td>
</tr>
<tr>
<td>Spanish</td>
<td>French</td>
</tr>
<tr>
<td>Estonian</td>
<td>Turkish</td>
</tr>
</tbody>
</table>

in-degree and eigenvector centralities.

Next, we performed a correlation analysis between language centralities and \( \text{HDI}_\ell \). For both, the G20 and the Olympics datasets, we found that the \( \text{HDI}_\ell \) positively correlates with three centralities. For the others, we can not derive interpretations as we could not find statistical evidence to support their correlation values. Table 5.6 summarizes the correlation results. We found that eigenvector, in-degree, and out-degree positively correlate with \( \text{HDI} \); the first two with confidence level of 99%, while the out-degree at 95%.

The out-degree centrality in a language network can be seen as a measure of multilingualism, i.e., the amount of other languages used by the speakers of a particular language. The in-degree, on the other hand, express the amount of interest of other languages in a particular language; it can be seen as a globalization indicator of a language. Last, the eigenvector centrality considers the structural connectivity of a language, not only measuring the diversity of connections, but also the volume (flow) potentially exchanged through these connections. The eigenvector centrality is not only a more robust metric, but it gave us the best correlation
Table 5.6: The correlation values between language centralities and HDI$\ell$ in our datasets. Only Eigenvector and In-degree centralities have positive correlation with HDI$\ell$ at 99% confidence level.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>G20</th>
<th>Olympics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eigenvector</td>
<td>0.46***</td>
<td>0.46***</td>
</tr>
<tr>
<td>In-degree</td>
<td>0.44***</td>
<td>0.45***</td>
</tr>
<tr>
<td>Out-degree</td>
<td>0.35**</td>
<td>0.31**</td>
</tr>
<tr>
<td>Closeness</td>
<td>0.34**</td>
<td>0.22</td>
</tr>
<tr>
<td>Betweenness</td>
<td>0.09</td>
<td>0.09</td>
</tr>
<tr>
<td>Weighted In-Degree</td>
<td>0.09</td>
<td>0.08</td>
</tr>
<tr>
<td>Weighted Out-Degree</td>
<td>0.09</td>
<td>0.19</td>
</tr>
</tbody>
</table>

Significance codes: *$\rho < 0.10$, **$\rho < 0.05$, and ***$\rho < 0.01$.

results with HDI$\ell$, our proxy for quality of life.

Figure 5.6 visually demonstrates the correlation between the HDI$\ell$ and their positions in the networks. In the plots, languages are represented as colored circles, in which color and size are visual clues for the standard deviation $\sigma$(HDI$\ell$) (Equation 5.5). A higher $\sigma$ means a language is spoken in countries with more different HDI$\ell$, i.e., heterogeneous countries. Most languages with high standard deviation are those spoken in imperialist countries, such as French, English, and Spanish. Also, those languages are always bellow the fitting correlation line, meaning these diversity of countries are leading to less quality of life than expected when comparing to languages’ centralities.

### 5.2.4 Discussion

In this simple application, we set to understand if we can relate the language connection patterns of users on Twitter to standard of living aspects in the real-world, such as HDI. We also observed that a few popular languages that have very high in-degree and eigenvector centrality do not tend to be the ones with very high
Figure 5.6: The significant positive correlation between language centralities and the Human Development Index indicate that the language positions in the network can be co-related to the QoL. Size and color of the languages represent the standard deviation of the estimated HDI.ₗ.
HDI. It is worth noting that some of the popular languages are spoken in different parts of the world. We demonstrate that overall the positions of languages correlate significantly with the $\text{HDI}_L$ (HDI of the languages).

Our work can be extended to understand the other factors that can be added along with the language positions to better describe the variability of the QoL. We also aim to analyze geo-tagged tweets to understand the current location of users and the relationship with the languages they choose to use.

It is important to note that the ranks of the number of speakers by language in our datasets, despite similar to each other, do not reflect the estimates of the world speakers by languages. Ethnologue reports that Chinese is the most spoken language followed by Spanish and English [60], but ours show the predominance of English followed by Spanish. There are several factors that may influence the different results: the penetration of Twitter in the population depends of age and census composition of the users. However, the disparities do not hinder in extracting interesting insights about the speakers in countries where Twitter is popularly used.
Chapter 6

Conclusions, Limitations, and Future Works

In this work, we proposed a data driven approach for information retrieval based on social interaction using social media data. The methodology consists in characterizing users, in relation to a set of entities, based on the amount of attention given to them. The characterization of entities and places emerge from several forms of aggregating users—a collective behavior. The unstructured data is standardized into characterization matrices. Thus, we propose the application of traditional techniques from statistics and machine learning as exploratory tools to visualize the data, to identify relationships, to highlight less prevalent behaviors, to rank entities, to cluster, and to obtain networks.

Throughout this dissertation, we tried to answer to what degree online social media data can be used to identify behaviors and relationships from social actors in a way to enable knowledge extrapolation. The results obtained by applying the proposed methodology in the three different domains support our hypothesis, i.e.,
(i) users can be characterized based on their social activity in relation to a set of entities—football clubs, languages, and organs; (ii) these characterizations can be used to extract knowledge about entities relationships and the interplay between entities and society—significant positive correlation with traditional football supporter ranks, language locality effect, and organs cascade failure effect; and (iii) the new knowledge is valid, and can be used to support further social analyses—significant proxy for social disorganization based on entropy of supporters in a region, and significant positive correlations between quality of life and languages’ importance. Figure 6.1 summarizes this work based on the frequency of words used in the dissertation.

We demonstrate that a straightforward user characterization of social media data allows us to grant knowledge access and provide insights to specialists, and demonstrate the richness of the enormously available data. We chose 3 domains to apply the proposed approach—sports (football), culture (language), and health (organ donation). The characteristics of these domains allow us to explore different aspects of our methodology and the richness of Twitter data, for instance:

**Generalization** – the diversity of domains increases the confidence about the generality of the proposed method, revealing the plurality of subjects embedded on social media.

![Figure 6.1: Word cloud formed by this whole dissertation.](image)
Size and locality of datasets – we used several datasets at city-, country-, and global-level.

Different forms of collecting Twitter data – we explored distinct sampling methods while collecting data, such as no filter, keywords, hashtags, mentions, and bounding boxes.

Number of entities in each domain – the characterizations ranged from a few to hundreds entities and places.

Dynamics of relationships between users and entities – the chosen subjects in our characterizations carry distinct engagement levels between users and entities. For instance, sport rivals or organ donation campaigns can be defined in a short-term periods; favorite clubs and secondary languages are more close related to long-term choices; mother tongues and transplanted organs are super-long-term (static) relationships.

In the football application, we showed traditional ranks of clubs popularity and unpopularity significantly correlate with the metrics we proposed to analyze social data. In addition, we were able to identify rivalries, and group clubs based on their similar rivals and rivalry intensity. In the language work, we noticed that Twitter datasets are very sensible to locality, i.e., we are capable of characterizing a language worldwide or in a particular city, and capture their singularities. Moreover, people tend to use languages based on neighborhood’s language more than based on language syntax proximity (language tree). Last, the characterization of organs captured dual transplantation dependencies and cascade failure effects over organs. Also, it unveiled distinct necessities among US states suggesting that generic allocation polices may not be the best approach in this subject.
Later, we demonstrated with two extended works how the knowledge generated by this exploratory analysis can be used for further investigations. First, we used diversity of football supporters in different regions of the UK to estimate levels of social disorganization by measuring criminal activities. Then, we showed the potential of language networks to indicate the quality of life (HDI) of the speakers of particular languages.

Finally, we expect these contributions, especially the proposed methodology, help the scientific community advance science. We deeply wish future research motivated by this dissertation to be driven by ethical values, promoting scientific knowledge via a better understanding of our society.

**Limitations**

Despite all these findings, our work has limitations. Since the general modeling aspect given by the use of our methodology, the type of questions we can answer are necessarily more shallow; similar to the trade-off between exploration and exploitation.

Not only the questions are limited, but also the themes (topics) capable of being investigated. By the public nature of social media, several topics are not properly discussed, thus our methods can not be applied to these neglected topics. Also, it is not enough to choose a subject to explore, one has to define a finite set of entities. For some problems, it would be interesting to capture unknown/new entities. For this cases, our approach might not be indicated, especially when the data collection is targeted based on the selected entities; *a posteriori* inclusion of entities may not capture the whole attention given to them. Moreover, the
methodology handles individuals democratically when trying to extract information from their relationships, i.e., like in an election process where every citizen has an equal weighted vote. The normalization phase during the user characterization may not be recommended when trying to understand influential dynamics among users/entities. In these scenarios, more active users could indeed have more power.

Our applications have limitations too. First, there is intrinsic bias in Twitter data, therefore our conclusions are limited by the scope of the social media site. More important, we are computer scientists developing tools to help other scientists by providing new types of data. So, sometimes we tend to cross the grey line between finding hidden patterns on data and interpret them. This ultimate understanding and explanation of human behavior should not be performed without the proper involvement of the specialists in the fields.

Finally, the proposed methodology presented here is not a toolbox in which one inputs social data to get publishable results. There is no such easy science. Probably, we are providing a simpler way to do the how’s, but the what’s and why’s are by no means alleviated.

**Future Works**

The methodology put forward can be improved by incorporating more data analysis, for instance, to explore time-series aspects of the characterizations, or to incorporate sentiment analysis in the characterization. With further applications, we might be able to be more specific in the last block of our method, i.e., how to use the knowledge generated by the characterizations as input for other research. We also plan to make it available via code to truly achieve our objective.
of disseminating the use of social media data in scientific studies.

We instantiated our framework in three domains, but there are a lot more to be explored. Moreover, we claim our methodology is valid for social media in general, but we only applied it using Twitter data. Twitter itself might be different due to changes on character count limits, therefore, we should investigate how/whether this increasing will impact on characterizations. Although we could not identify incompatibilities between the myriad of social media sites compromising the characterization framework, we should try to apply this framework using different sources of data.

Finally, after some experimental results, we believe this methodology can be generalized to other types of data rather than social media. Maybe, this indirect way of learning about entities by relatively comparing different perspectives is something more related to how we human experience things in broader scenarios.
References


Ransen Niu, Jiawei Zhang, and David S Ebert. Classification and Visualization of Crime-Related Tweets. In The Summer Undergraduate Research Fellowship (SURF) Symposium, 2015.


[129] Lara Rebello. UN Human Development Index Report: Norway leads for the 12th year; UK comes in 14th, 2015.


Appendix A

Validation

In this appendix, we describe how the validation process proposed on Section 3.2.3 of the methodology was conducted throughout this work. Despite being particularly related to the applications presented on Chapter 4, we decided to report the validation results independently for the following:

**Scope** – the validation presented here goes beyond the assessment of the characterizations on Chapter 4; we investigate the concepts applied, i.e., if the subjects and entities are meaningful. For instance, we use larger datasets than the ones used on Chapter 4. We also evaluate longitudinal implications, something not addressed by our applications.

**Repetition** – we applied the same method in all applications – football, language, and organs – generating similar figures. Therefore, repeatedly reporting them after each characterization would result in very prolix text.

**Flow** – alternatively to excessive repetition, we could report the results compressed together. Thus, this validation report could be placed as the last
We proposed a two-level validation, first checking the specificities related to users, and later, evaluating the aggregations. In general terms, we evaluate: (i) types of users and their implications; (ii) singularities; (iii) temporal fluctuations; (iv) rank stability; and (v) sample size. Table A.1 present some statistics from the four datasets used in this validation. They cover different aspects interesting to this validation with a wide range of statistics, such as – number of days ($10^1$–$10^2$), entities ($10^0$–$10^2$), users ($10^4$–$10^6$), tweets ($10^5$–$10^7$).

Table A.1: Twitter datasets used in the validation process.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Football BRA</th>
<th>Football UK</th>
<th>Language</th>
<th>Organs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start Data Collection</td>
<td>06/May/16</td>
<td>07/Feb/15</td>
<td>01/Ago/16</td>
<td>18/Apr/15</td>
</tr>
<tr>
<td>Finish Data Collection</td>
<td>11/Apr/17</td>
<td>07/May/15</td>
<td>24/Ago/16</td>
<td>11/May/16</td>
</tr>
<tr>
<td>Number of Days</td>
<td>339</td>
<td>89</td>
<td>22</td>
<td>388</td>
</tr>
<tr>
<td>Tweets</td>
<td>26,119,376</td>
<td>4,920,764</td>
<td>2,630,412</td>
<td>134,265</td>
</tr>
<tr>
<td>Users</td>
<td>1,066,550</td>
<td>1,151,702</td>
<td>21,370</td>
<td>71,947</td>
</tr>
<tr>
<td>Entities</td>
<td>330</td>
<td>20</td>
<td>53</td>
<td>6</td>
</tr>
</tbody>
</table>

A.1 User Validation

When dealing with real data, especially, social network data, it is expected to find very distinct usage patterns, i.e., there are very active, but also quiet users. Because of that, the proposed methodology focused on individuals rather than social activities (posts) to reduce possible bias towards the most active users.
Nonetheless, we need to verify which other aspects may differentiate those users, and how they could affect their characterization.

To handle with these varieties of users, we divided them in 10 color coded groups based on their activity level. For validation purposes, in order to capture temporal fluctuations, we only considered users with 50+ tweets in the user validation. Figure A.1 shows the number of users per group in the 4 datasets used in this validation chapter – Football in Brazil, Football in the UK, Languages in Olympics, and Organs in the USA. As the average number of tweets per user in the organ dataset is lower, the constraint of 50+ tweets per user applied in this validation

![Diagram](a) Football BRA  ![Diagram](b) Football UK  ![Diagram](c) Languages  ![Diagram](d) Organs

Figure A.1: Number of users per activity-level group. For each dataset, users are split into 10 groups with logarithmic bins. The minimum activity level considered was 50 tweets. Since the datasets capture distinct subjects and encompass variable time-span, the range of tweets per group varies among them. For instance, the range in first group (purple) is: 50-90 in the Football BRA dataset (a), but 50-69 otherwise.
reduces drastically our sample size. Consequently, the results are less conclusive for this dataset as depicted by larger confidence intervals.

As seen in the inner plots of Figure A.2, the number of tweets per user indeed follows a fat-tailed distribution, justifying the exponential bins separating users of Figure A.1. Knowing users are intrinsic distinct, at least by the amount of online interaction measured by the number of posts, we investigated how this could affect the number of entities being mentioned (Figure A.2). As expected, the more users tweet, they are more likely to mention more clubs. However, all types of user seem to mention new entities at the same rate. Overall, each user only mentions a small percentage of all entities.

Since users tend to mention new entities with time, a natural question to raise is how the characterizations evolve with time and how we could compare these characterizations. A simple use of the characterizations is to transform the frequency of attention given to each entity in a rank of preferred entities. So, the preferred entity is the one most mentioned, the second in the rank is the second

![Figure A.2: Average number of entities mentioned by users. Plots from different datasets showing the cumulative evolution of unique entities being mentioned. Solid lines represent the mean of 500 random sampled users (when applicable) colored by activity-group, and shaded areas are the respective 95% CI. Inset plots showing the distribution of tweets per user (blue dots), with dashed-fitting lines for power-law (orange) and log-normal (green) distributions.](image-url)
most mentioned, and so on. How mutable is this rank of a user from his first tweet until the last one? In the best case scenario, each time a user mention a new entity the characterization would place it correctly in rank.

Table A.2 shows a toy example. At time $t_1$, user mentions his first entity $A$; thus, it is his preferred entity (1$^{st}$ in the rank) while all others entities, not mentioned yet, are in the last position in the rank (i.e., 2$^{nd}$). At time $t_2$, user mentions 2 entities, $B$ for the first time and $A$ again; cumulatively, $A$ is still his preferred entity$^1$, $B$ becomes his 2$^{nd}$ preferred, and the rest are in 3$^{rd}$; and so on and so forth. Nevertheless, the rank of entities wouldn’t be affected. Figure A.3 shows the correlation evolution results for an example where users mention from 3 to 30 of 300 entities in the best possible sequence. We cannot establish a value as a good correlation threshold since they vary substantially depending on the number of mentioned entities. Figure A.4 shows this temporal evolution measured for our datasets. Plots are very similar to the hypothetical one, suggesting users may be

---

$^1$According to the methodology (Chapter 3), multiple mentions in a single post should be weighted accordingly, but for didactic proposes Table A.2 is not weighting the divided attention.

Table A.2: Toy example of the best case mentioning scenario of a user mentioning 4 of 10 entities – $ABCDEFGHIJ$. Entities not mentioned are assigned to last position in the rank. For each time $t_i$, Spearman correlation $\rho$ is calculated between mentioned rank in $t_i$ and the final rank $t_4$.

<table>
<thead>
<tr>
<th>Time</th>
<th>Tweet</th>
<th>Count</th>
<th>Rank</th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_1$</td>
<td>“Love $A$”</td>
<td>1000000000</td>
<td>1222222222</td>
<td>.588</td>
</tr>
<tr>
<td>$t_2$</td>
<td>“Love $AB$”</td>
<td>2100000000</td>
<td>1233333333</td>
<td>.789</td>
</tr>
<tr>
<td>$t_3$</td>
<td>“Love $ABC$”</td>
<td>3210000000</td>
<td>1234444444</td>
<td>.916</td>
</tr>
<tr>
<td>$t_4$</td>
<td>“Love $ABCD$”</td>
<td>4321000000</td>
<td>1234555555</td>
<td>1</td>
</tr>
</tbody>
</table>

163
Figure A.3: Temporal evolution of hypothetical characterizations based on Spearman correlation. Best case evolving scenarios for a 300-entity characterization for individuals mentioning from 3 (blue line) to 30 (brown line) entities. At each new tweet, a new entity is mentioned and the entity’s rank obtained from its characterization matches the final true one.

Figure A.4: Temporal evolution of user characterizations based on Spearman correlation. Plots comparing the final characterizations (all tweets) to previous ones at each new tweet. Solid-lines are the mean self comparison, i.e., the mean correlation between the same user at different times. Groups are represented by 500 random sampled users (when applicable), and shaded areas are the respective 95% CI.

Figure A.5, on the other hand, shows the same temporal evolution of user characterizations, but measuring the similarity via Bhattacharyya distance instead of Spearman correlation. This time, the distinction between users of different mentioning profiles are much more similar. Moreover, Figure A.4 shows not only the self-comparison evolution, but how similar a user is from others in the same coherent along time with their preferences.

Figure A.5, on the other hand, shows the same temporal evolution of user characterizations, but measuring the similarity via Bhattacharyya distance instead of Spearman correlation. This time, the distinction between users of different mentioning profiles are much more similar. Moreover, Figure A.4 shows not only the self-comparison evolution, but how similar a user is from others in the same
Figure A.5: Temporal evolution of characterization based on Bhattacharyya distance. Plots comparing the final characterizations (all tweets) to previous ones at each new tweet. Solid-lines are the mean self comparison, i.e., the mean distance between the same user at different times. Dashed-lines are the mean distance between a user and all others in the same group. Dotted-lines are the standard deviation of the mean dashed-lines. Since the first tweet, a user is much more similar to his/her final characterization than to any other user. Groups are represented by 500 random sampled users (when applicable), and shaded areas are the respective 95% CI.

Another important aspect to investigate is rank stability. Again, a natural way to use and interpret the characterizations are by ranking entities based on their mentioned score. Before addressing stability, we need to answer other question – How mentioned scores are affected when more entities are being mentioned? Figure A.6 shows the distribution of mentioned scores for top-5 rank positions in the four datasets being studied. Regardless the number of tweets, and therefore the number of mentioned entities (Figure A.2), each rank position has its typical value for each dataset.

Knowing different groups of users have similar mentioned scores is good because
Figure A.6: Distribution of mentioned score (Equation 3.3) for the top-5 rank positions for different groups of user. The peaks are barely indistinguishable meaning they assume similar values regardless users’ activity level.

Figure A.7: Distribution of mentioned score for the top-5 most mentioned entities for the group of users with 292-527 tweets of the Brazilian football dataset. In the left panel, we can clearly see rank 1 (red) apart from others. In the right panel, zooming in the lower ranks to show only the peak of rank 2 is distinguishable from others.
we do not need to develop group-specific methods nor interpretation of the results. However, is there a typical mentioned score for each rank position? How likely is to users to change their preferred entity? What about their second preferred one? Figure A.7 shows the distribution of mentioned scores for the top-5 ranked entities for the group of users tweeting in the range 292-527 in the Brazilian football dataset. From the left panel we can see a clear distinction from rank 1; zooming in the right panel, we see rank 2 having a spread behavior while the other have coincidental peaks.

Figures A.8–A.11 present a visual evolution of the top-5 ranked entities for each investigated dataset. Each figure have five plots, one for each top-5 positions, showing the median rank at each new tweet of the final entity at that position. For instance, given $B$ is the second most mentioned entity of a user when he/she is characterized using all his/her $t$ tweets, then we get $B$’s position in all $t-1$ characterizations to represent as the second rank evolution. These plots are showing the median evolution for each rank position for each group of users. For most datasets and up to rank 3, entities are correctly ranked as earlier as 5% of data is processed.

A more systematic way to check whether the proposed characterization leads to stable ranked entities is to compare the difference of the mentioned score of consecutive ranked entities ($\Delta$, Equation 3.9) against the fluctuations around the mean mentioned score ($\delta$, Equation 3.10) as described in Section 3.2.3. Figure A.12 shows the analytical evidence for stable ranked entities ($\Delta > \delta$). In all datasets, the preferred entity and the second most mentioned are stable. The former is clearly super stable while the latters stability become more evident for users with higher number of tweets. When comparing the different applications,
Figure A.8: Football BRA – Temporal median rank evolution of the final top-5 preferred entities per individual. The preferred entity (rank 1) is always correct. Up to rank 3, most group of users correctly rank entities with less than 5% of data. Horizontal dashed red lines showing $\pm 0.5$ of the respective rank, while horizontal dashed black lines show the rank expected final position. Vertical dashed lines marking the position where 5% of tweets for each group. Groups are represented by 500 random sampled users (when applicable).
Figure A.9: Football UK – Temporal median rank evolution of the final top-5 preferred entities per individual. The preferred and second entities (rank 1 and 2) are always correct. For rank 3, most group of users correctly rank entities with less than 5% of data. Horizontal dashed red lines showing ±0.5 of the respective rank, while horizontal dashed black lines show the rank expected final position. Vertical dashed lines marking the position where 5% of tweets for each group. Groups are represented by 500 random sampled users (when applicable).
Figure A.10: Languages – Temporal median rank evolution of the final top-5 preferred entities per individual. The preferred, second, and third entities (rank 1, 2, and 3) are always correct. Horizontal dashed red lines showing ±0.5 of the respective rank, while horizontal dashed black lines show the rank expected final position. Vertical dashed lines marking the position where 5% of tweets for each group. Groups are represented by 500 random sampled users (when applicable).
Figure A.11: Organs – Temporal median rank evolution of the final top-5 preferred entities per individual. Only the preferred and second entities (rank 1 and 2) are always correct. Horizontal dashed red lines showing ±0.5 of the respective rank, while horizontal dashed black lines show the rank expected final position. Vertical dashed lines marking the position where 5% of tweets for each group.
Figure A.12: Analytical evidence for stable preferred entities. Mentioned score gap $\Delta$ and the fluctuation $\delta$ plotted as function of the rank for different groups of users and datasets. Super stability for the preferred entity in all possible scenarios. The second-preferred entity is also stable (dashed-red line) for all groups, but less evident. As users become more active, they have more stable ranked entities.
organ characterization seems to provide more stable ranking positions followed by language and football.

A.2 Aggregated Validation

In addition to validation of individual characterization, we need to address similar issues when we have an aggregated characterization, specifically the preferred entity aggregation.

The objective of aggregated characterizations is to capture the emergent behavior of its representing users; in this case, users sharing their preferred entity. A typical question to address when designing data driven application is the amount of data that is necessary to provide confident results. So, we tested the aggregated rank stability under different samples size of users. For instance, when using all 10,000 users, entities are ranked as ABCDE. When sampling 10, 100, or 1,000, if the ranks are identical, we say this characterization is very stable (at least, the top-5 ranked positions). However, if the sampled ranks are AXTBC, ABXTC, and ABTCX, we would say only the first two positions are stable.

Figure A.13 shows the impacts of user sampling on the final rank of the aggregated characterizations. For each dataset, we consider as the truth rank the characterization achieved with all users. Then, for each sample, we performed the characterization, and listed the respective rank position for the true fifth first ones. The box-plots showed in Figure A.13 represents the positions obtained by resampling 10 times. The correct rank is quickly reached; the median top-5 rank is corrected obtained with only 0.9%, 0.1%, 4.7%, and 0.7% of the users from football BRA and UK, languages, and organs datasets, respectively.
Figure A.13: High top-5 rank stability even for tiny user samples. Box-plots represent bootstrapping with 10 resamples of X users. Using the sampled users we characterize entities and compare the top-5 rank against the full characterization. The median top-5 rank is corrected obtained with only 0.9%, 0.1%, 4.7%, and 0.7% of the users from football BRA and UK, languages, and organs, respectively.
By the sampling analysis, one may conclude that collecting a lot of data is useless, or even counter productive. However, despite being precise with very few users, Figure A.14 shows how the coverage are affected, i.e., with few users we can get good characterization, but only few of them. This issue is more evident when the set of entities is large, and users have very unbalanced prevalence of preferring them. Many times, we researchers have limited access to social media data. Twitter, for instance, usually offers 1% publicly; only minorities have access to 10% or more. From this sampling investigation, we can conclude this limited data (volume) does not give misleading results, at most, narrow ones.

Another way to test the rank stability is to check how they evolve in time. This might be helpful in deciding how long to collect data. Having 1 billion tweets in a single day is complete different of having 1 million tweets in 1 thousand different days. The volume of data is important as we already showed in figures A.13 and A.14, but the time-window is crucial for longitudinal analysis. Figures A.15 and A.16 present the temporal evolution of the top-5 ranked entities in our datasets. They also show the impact of the time-window in the coverage. Unlike sampling users, smaller time-windows have less impact on coverage, suggesting the diversity of users on a daily base.

Yet, we can observe the rank stabilization not as an universal property of the aggregated characterization, but dependent of context. For instance, language characterizations seem to be more stable than football clubs. Figure A.16 captures two interesting aspects involving the Brazilian football. First, the tragic accident involving Chapecoense in which almost all its players have lost their lives. This terrible event reverberated worldwide, touching supporters and changing their secondary preferences. The second aspect is how the Brazilian football is scheduled
Figure A.14: Coverage based on different sample sizes of users. Average percentage number of entities characterized measured by bootstrapping 10 resamples of X users. For instance, with only 10% of users we are able to characterize 50%, 100%, 50%, and 80% of entities.

over the year. From May to December, they have national (series A, B, C, and D, and Brazil’s cup) and international (Libertadores and Sul-Americana) competitions; from January to April, they have the 27 state’s leagues. Even with the huge number of national and international competitions, many clubs (around 20%) only have audience during the state’s tournaments.

Finally, we perform a similar formal rank stability test as we did for users. Figure A.17 demonstrates that aggregated characterizations represent not only the emergent behavior of their users, but they are collective much more stable
than their composing users.

Figure A.15: Temporal analyses of aggregated characterizations. On the left panel, the mean rank evolution of the final top-5 preferred entities per entity. The stabilization for the English football takes longer. On the right panel, the percentage of entities characterized per rank position (solid lines) and the rapid decay of the Bhattacharyya distance when comparing characterizations day-by-day (dashed line).
Figure A.16: Temporal analyses of aggregated characterizations using the Brazilian football dataset. On the left panel, the mean rank evolution of the final top-5 preferred entities per entity. On the right panel, the percentage of entities characterized per rank position (solid lines) and the rapid decay of the Bhattacharyya distance when comparing characterizations day-by-day (dashed line). The black vertical dashed line marks Nov 28\textsuperscript{th}, the fatal day of the airplane crash involving the Chapecoense club killing almost his entire players. On (b), we perform the analyses removing Chapecoense from the dataset.
Figure A.17: Aggregated characterizations are more stable than individual characterization (compare with Figure A.12). Analytical evidence for stable preferred relationships on entities characterization. Mentioned score gap $\Delta$ and the fluctuation $\delta$ plotted as function of the rank for different datasets. Super stability for the top5 ranked entities in all possible scenarios. Inner plot in (a) showing the rank stability for the Brazilian football is equivalent when Chapecoense is removed from the dataset.
Appendix B

Supplementary Figures

In this section the reader may find supplementary figures supporting the results shown in the text. All figures are self-contained, with their explanation in the caption.
Figure B.1: Engagement on Brazilian football dataset from May-16 until Mar-17. The effects of the tragic plane crash involving the club Chapecoense on data collection. (a–b) Engagement of users (circle) and entities–clubs (triangle) per week. Top panel showing aggregated results per week, while bottom panel shows cumulative figures. Colors used to separate users into groups restricting the minimum number of tweets, from purple (at least 1 tweet) to red (minimum 100 tweets). On (a), all data reveal a huge spike on the week of the accident, while on (b) the week of the crash was removed. (c) Shows the spike of tweets was mentioning Chapecoense. (d) The day of the accident impacted the distribution of tweets per week, compare the red and black lines.
Figure B.2: Ranking the BSA supporters. On top, clubs are ranked based on percentage of users (red) and number of tweets (blue). On bottom, clubs are ranked considering only the mentioned score of their preferred users; a possible proxy for passionate level or purity of supporters.
Figure B.3: Ranking the EPL supporters. On top, clubs are ranked based on percentage of users (red) and number of tweets (blue). On bottom, clubs are ranked considering only the mentioned score of their preferred users; a possible proxy for passionate level or purity of supporters.
Figure B.4: Number of users per language in different datasets calculated using Equation 3.11.
Figure B.5: Language characterizations from the Olympics16 dataset. Y-axis represent probabilities ($k_{jw}$) of the amount of attention given to a language in log scale and the three most used languages other than itself are highlighted.
Figure B.6: Language characterizations from the Tokyo dataset. Y-axis represent probabilities ($k_{ijw}$) of the amount of attention given to a language in log scale and the three most used languages other than itself are highlighted.
Figure B.7: Language characterizations from the New York City dataset. Y-axis represent probabilities ($k_{ju}$) of the amount of attention given to a language in log scale and the three most used languages other than itself are highlighted.
Figure B.8: Language characterizations from the Paris dataset. Y-axis represent probabilities ($k_{jw}$) of the amount of attention given to a language in log scale and the three most used languages other than itself are highlighted.
Figure B.9: Relative risk (log) of organs for each US state. Y-axis show the organs – heart (1), intestine (2), kidney (3), liver (4), lung (5), and pancreas (6). Organs are significantly mentioned if their confidence interval do not cross zero (vertical red line) – less than expected in blue and more than expected in red. Organs are labeled with their percentage in the state if RR is significant and there are more than 5 users mentioning them.
Appendix C

Twitter Datasets

In this section we present the collection process to generate the Twitter datasets presented on Table 4.5. Datasets are listed from more global to more local and labeled global-level°, country-level†, and city-level*: 

2016 Olympic Games° – It refers to the 2016 Olympic Games hosted in Brazil with more than 11,000 athletes from 207 countries. We tracked the 64 different translations for the word “olympics”. We used the Google Translator to get translations from 107 languages. Figure C.1 shows the list of tracked (translated) terms.

Figure C.1: The term “olympics” translated to 107 languages according to Google Translate. These are the tracking terms to build the Olympics16 dataset.
G-20° – We tracked the last name of the leaders of the Group of Twenty (G-20) in 2014: Zuma, Kirchner, Roussef, Pena Nieto, Harper, Obama, Jinping, Noda, Myung-bak, Singh, Yudhoyono, Abdullah, Erdogan, Merkel, Hollande, Renzi, Putin, Cameron, and Gillard.

2015 Womens’ World Cup & America Cup° – It refers to these two concurrent football competitions in 2015, the former hosted in Canada and the later hosted in Chile. We tracked the 3-letter hashflags and Twitter official accounts from the 31 competing national squads. Brazil, USA, China, Nigeria, Ecuador, Canada, Thailand, France, England, Cameroon, Côte d’Ivoire, Switzerland, Spain, Colombia, Costa Rica, Sweden, Norway, Mexico, Australia, Japan, Korea, New Zeland, Germany, Argentina, Bolivia, Chile, Jamaica, Paraguay, Peru, Uruguay, and Venezuela.

2014 FIFA World Cup° – It refers to the most popular sport event in the world, hosted in Brazil. We tracked the 32 nation squads’ 3-letter hashflags. Algeria, Cameroon, Côte d’Ivoire, Ghana, Nigeria, Australia, Iran, Japan, Korea Republic, Belgium, Bosnia and Herzegovina, Croatia, England, France, Germany, Greece, Italy, Netherlands, Portugal, Russia, Spain, Switzerland, Costa Rica, Honduras, Mexico, USA, Argentina, Brazil, Chile, Colombia, Ecuador, and Uruguay.

2016 UEFA Euro° – It refers to the football event hosted in France. We tracked the 32 nation squads’ 3-letter hashflags. Albania, Austria, Belgium, Croatia, Czech Republic, England, France, Germany, Hungary, Iceland, Italy, Northern Ireland, Poland, Portugal, Republic of Ireland, Romania, Russia, Slovakia, Spain, Sweden, Switzerland, Turkey, Ukraine, and Wales.
The United Kingdom† – These are tweets collected using the bounding box (BB) of the United Kingdom, which encompasses four countries: England, Scotland, Wales, and Northern Ireland.

South America† – These are tweets collected using the BB of Brazil, which encompasses most areas and countries of South America, except Ecuador.

New York City*, Paris*, San Francisco*, Tokyo*, and Hong Kong* – These are geo-tagged tweets within the bounding box of the city.