Characterization of the Twitter @replies Network: Are User Ties Social or Topical?

Daniel Sousa
DEI-FEUP, University of Porto
Rua Dr. Roberto Frias, s/n
4200-465 Porto, Portugal
djrsousa@gmail.com

Luís Sarmento
Labs SAPO and LIACC/FEUP
Rua Dr. Roberto Frias, s/n
4200-465 Porto, Portugal
las@co.sapo.pt

Eduarda Mendes Rodrigues
DEI-FEUP, University of Porto
Rua Dr. Roberto Frias, s/n
4200-465 Porto, Portugal
eduardamr@acm.org

ABSTRACT

In recent years, social media services have become a global phenomenon on the Internet. The popularity of these services provides an opportunity to study the characteristics of online social networks and the communities that emerge in them. This paper presents an analysis of the users' interactions in the implicit network derived from tweet replies of a specific dataset obtained from a popular microblogging service, Twitter[1]. We analyze the influence of the topics of the tweet messages on the interaction among users, to determine if the social aspect prevails over the topic in the moment of interaction. Thus, the main goal of this paper is to investigate if people selectively choose whom to reply to based on the topic or, otherwise, if they reply to anyone about anything. We found that the social aspect predominantly conditions users' interactions. For users with larger and denser ego-centric networks, we observed a slight tendency for separating their connections depending on the topics discussed.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Systems—*Web-based services*; J.4 [**Computer Applications**]: Social and Behavioral Sciences—*Sociology*

General Terms

Experimentation, Human Factors, Measurement

Keywords

Twitter, microblogging, social network, topic analysis

1. INTRODUCTION

The many social media services that have emerged over the past few years allow people to create virtual communities and to interact with each other online. The dynamics of social networks and the characterization of social ties is a topic of great interest in the data mining research community. Understanding social networks and user interactions is important to improve current systems and to design new applications, including personalized recommender systems (e.g., recommendation of *news* [12] or "friends"). Most social networks require users to make *explicit* social connections by accepting a "connection request". However, we can also consider *implicit* connections derived from users' activities in the social network, such as tagging another user's photo, sending a message, or posting a comment on other user's page. In some ways implicit networks can be better indicators of the actual social ties between users. According to a study of Facebook [5], we know that people only communicate with a few of their declared friends. Thus, one way of quantifying how strong a social tie is, is measuring the frequency of interaction between the two people involved [4].

In this paper, we focus on studying one type of implicit networks that can be found in Twitter, a popular microblogging service that emerged in 2006 [6]. The informal and spontaneous nature of this social media service encourages users to interact frequently with potentially large communities. Users may choose to follow other users, which enables them to receive all the tweets from the followed users. Unlike the messages posted to topic-specific mailing lists, the tweets may cover a very wide range of topics. Users may *reply* to any specific tweet, thus engaging in a kind of conversation. We aim to analyze the influence of the topic in users' decision to reply to tweet messages. More specifically, we aim to investigate if users are motivated to reply based on social ties, thus communicating with their "friends" about *any* topic or if, on the other hand, they discriminate who they reply to based on the topic of the original message.

The answer to this question is particularly important for designing systems that recommend new social connections (i.e., new "friends"). If users do actually distinguish their connections based on the topics, then a *content-based* recommender system might be useful for suggesting interesting connections. Otherwise, friend recommendations should be essentially based on the social network structure, such as the *People You May Know (PYMK)* system used in MySpace [11], rather than on the content of the messages exchanged.

This paper is organized as follows. First, we review related work and then we explain the adopted methodology that allows us to investigate if users distinguish their connections according to the topics discussed. Next, we present

[1]http://www.twitter.com

SMUC'10, October 30, 2010, Toronto, Ontario, Canada.

a brief description of the dataset of Twitter messages and associated replies, which we used for deriving an implicit social network. Then, we describe our experiments and present our results. Finally, we discuss our findings regarding the influence of the topic on users' interactions, and we conclude by outlining areas for future work.

2. RELATED WORK

In recent years, several studies have been conducted on social networks. Many of them have discussed extensively topological and structural issues of the networks, including their size, density, degree distributions, geodesic paths, connected components, community structure, among others [8, 9, 10, 13, 15]. However, an important aspect of social media services relates to the *activity networks*, which are networks that can be derived from the actual interactions between users, rather than from the mere declared friendship connections [4]. Viswanath et al. [14] built an implicit network based on Facebook wall posts, in order to study the evolution of activity among users. They found that the links in the activity network tend to come and go quickly, suggesting that the activity network is rapidly changing over time. Despite this high churn in users' interactions over time, they found that many of the global structural properties remain constant.

Early studies on Twitter include a systematic description of the main characteristics of microblogging and the impact on informal communication at work provided by this kind of services [17] and also try to understand how and why people use Twitter [6, 17]. Such studies reported a variety of social purposes and motivations to use Twitter, including to keep in touch with friends, to raise visibility of interesting things, to gather useful information according user's personal interests, or to get help and opinions on a specific matter.

A large-scale study of the conversational characteristics of Twitter and its power as a new medium of information sharing [9], reveals that any retweeted tweet is to reach an average of 1000 users regardless of the number of followers the original tweet has, showing the power of this service as a source of information dissemination. Boyd et al. [2] analyzed the conversational aspects of the retweets. The retweet mechanism empowers users to spread information of their choice beyond the reach of the original tweet's followers. Using a series of case studies and empirical data, Boyd et al. describe and map out the various conventions of retweeting, and examine retweeting practices. For example, they found that over 9% of all retweets include a reference to the retweeter's handle. In other words, A retweets B when B's message refers to A. They call these "ego retweets".

Java et al. [6] present an analysis of Twitter's growth rate, both in users and user posts, and provide insights into the user activity and geographical distribution. In general, they found that the reciprocity and clustering coefficient values for users in Europe and Asia are higher than the values obtained for users in North America.

Weng et al. [16] carried out an important study focused on the problem of identifying influential users on Twitter, and propose a measure that goes beyond the mere number of followers a user has. The measure is an extension of PageRank that measures the users' influence taking into account both the topical similarity between users and their network links' structure. Moreover, they also reveal that the presence of reciprocity in the explicitly following/follower network can be explained by the phenomenon of *homophily*. In other words, a user follows a *friend* because they have similar topic interests.

Summarizing, there are several research studies on Twitter and, to the best of our knowledge, none of them looked at characterizing topics' influence in the social network activity. In this paper we analyze the users' motivations for replying to other Twitter users trying to understand if users segment their links according to the topics discussed. This is important both to improve current systems and to design new applications (e.g., recommender systems).

3. PROBLEM SETTING

3.1 Twitter @replies Network

Twitter is a social networking and microblogging service that allows users to answer the question: "What are you doing now?". Twitter is also a social messaging service where users interact by sending short text messages with a maximum of 140 characters in length, called "tweets". Users can also follow other people and have friends, or "followers". Unlike many social media services, such as Facebook[2] or MySpace[3], the relationship of *following* and *being followed* requires no reciprocation. A user can follow any other user, and the user being followed needs not to follow back. Furthermore the practice of *responding* to a tweet has evolved into a well-defined markup vocabulary: RT stands for *retweet*, a '#' followed by a word represents an *hashtag*, and a '@' followed by a username directly addresses the message to that user, while keeping the message public for other users to read. This type of messages are often referred to as @replies. Therefore, users can be connected in an implicit or explicit manner as, for example, via retweets, #hashtags, @replies or through friends and followers relationships.

Considering the @replies messages, we can build an *activity network* representing the flow of interaction among Twitter users. Such implicit network derived from tweet replies can be represented as a directed graph $G = (V, E)$ where each node $u \in V$ represents a user and each edge $(u_i, u_j) \in E$ represents an @reply message sent from user u_i to the user u_j.

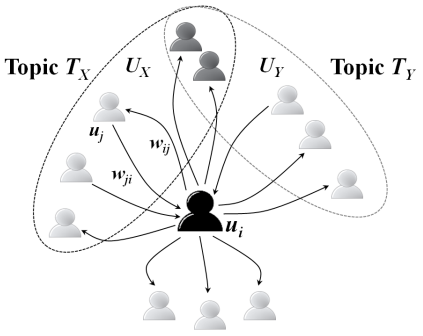

Figure 1: Ego-centric network of user u_i, representing @replies interactions with people who tweet on topic T_X (set of users U_X), topic T_Y (set of users U_Y), and other topics.

[2]http://www.facebook.com/
[3]http://www.myspace.com/

Every edge in the graph G also has an associated weight w_{ij} which corresponds to the number of replies sent from user u_i to user u_j, as depicted in Figure 1. This network can be analyzed from the perspective of each user u_i, by considering the individual's personal network or *ego-centric network* $G'(u_i)$, which comprises all the people to whom u_i sends at least one @reply message.

3.2 Topic vs. Social Motivations

This paper aims to understand what motivates a user to reply to other users tweets. More precisely, we seek to answer the following research question: "Are people interactions motivated by the *topic* or does the social aspect take precedence over the topic, when replying to tweets?". To answer this question we start by considering the ego-centric network of each user u_i, $G'(u_i)$. Given two topics, T_X and T_Y, we can extract from $G'(u_i)$ two, possibly overlapping subgraphs, $G'_{T_X}(u_i)$ and $G'_{T_Y}(u_i)$, which represent user u_i's replies to messages from other individuals on those two topics, respectively. Let U_X be the set of users to whom u_i is connected in $G'_{T_X}(u_i)$ and U_Y the set of users to whom u_i is connected in $G'_{T_Y}(u_i)$. There are two possible scenarios for these two sets of users:

1. There is no overlap between U_X and U_Y (i.e., $U_X \cap U_Y = \emptyset$);

2. There is a certain overlap between U_X and U_Y (i.e., $U_X \cap U_Y \neq \emptyset$). In some cases, the two sets can overlap completely. More precisely, one set of users can be a subset of the other, or the two sets can be exactly the same (i.e., $U_X \subseteq U_Y \vee U_X \supseteq U_Y$).

If user u_i's interactions are neither socially nor topically motivated, then u_i will not discriminate who he replies to based on any (of these) criteria. This is equivalent to saying that user u_i chooses whether to reply or not as if he was following a random process. Thus, let $p(U_X)$ and $p(U_Y)$ be the marginal probabilities of a user u_i interacting with other users on topics T_X and T_Y, respectively. Assuming no other selection criteria is being followed by user u_i, the probability of replying to a user in U_X, $p(U_X)$, should be independent of the probability of replying to a user in U_Y, $p(U_Y)$. Therefore, the probability of interacting with the same subset of people on both topics, which is equivalent to the probability of overlap between users in U_X and U_Y, will be given by $p(U_X,U_Y) = p(U_X) \cdot p(U_Y)$.

However, we know that users do not choose who they reply to randomly. So, we can measure how much $p(U_X,U_Y)$ differs from its expected value under the independence assumption and, thus, test whether user replies are motivated by the topic or by the social aspect. There are two possible cases:

1. If the user's interactions are topically motivated, then the actual observed probability of overlap will be inferior to the expected overlap. In such case, $p(U_X,U_Y) < p(U_X) \cdot (U_Y)$;

2. If, on the other hand, there is a social motivation for selecting who to reply to, the observed probability of overlap will be higher than expected under the independence assumption. Thus, $p(U_X,U_Y) > p(U_X) \cdot (U_Y)$.

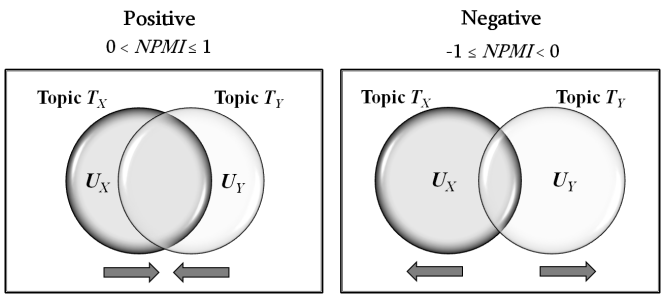

Figure 2: Interpretation of the *NPMI* values, considering the overlap between the sets of users U_X and U_Y.

3.3 Measuring User Motivation

To quantify the extent of the overlap and a user's motivation to reply to other users' tweets, we use the Normalized Pointwise Mutual Information (NPMI) [3], a bi-directional association measure used in information theory and statistics. This measure relates the probability of co-occurrence of two events with the probabilities of occurrence of each of them individually. In our problem setting, such events consist of user u_i's interactions with other users on topics T_X and T_Y, respectively. Thus, for a given pair of topics T_X and T_Y and associated sets of users U_X and U_Y, the NPMI measure is defined as:

$$NPMI(U_X,U_Y) = \frac{ln(\frac{p(U_X,U_Y)}{p(U_X)p(U_Y)})}{-ln(p(U_X,U_Y))}.$$

If the independence assumption holds, $p(U_X,U_Y) = p(U_X) \cdot (U_Y) \Rightarrow NPMI = 0$. In other cases, the NPMI association measure quantifies if the overlap between the two sets of users is higher than expected, i.e., $0 <$ NPMI ≤ 1, or, on the other hand, if it is lower than expected, i.e., $-1 \leq NPMI < 0$ (see Figure 2).

If U_X and U_Y are mutually exclusive, this implies that user u_i completely separates people in his ego-centric network according to the topics discussed. Thus, $p(U_X,U_Y) = 0 \Rightarrow NPMI = -1$. On the other hand, if U_X and U_Y consist of the same set of people, then $p(U_X,U_Y) = 1 \Rightarrow NPMI = 1$.

All the probability distributions at stake can be estimated using directly observable quantities (see Figure 3). In fact, the NPMI measure can be expressed based on such quantities. Let us define:

- $N = |G'(u_i)|$-1, as the size of u_i's ego-centric network, excluding u_i, i.e., the total number of users to whom u_i replied (in any topic);

- $a = \frac{|U_X \cap U_Y|}{N}$, as the fraction of users from $G'(u_i)$ to whom user u_i replied on both topics T_X and T_Y;

- $b = \frac{|U_X - U_Y|}{N}$, as the fraction of users from $G'(u_i)$ to whom user u_i replied only on topic T_X;

- $c = \frac{|U_Y - U_X|}{N}$, as the fraction of users from $G'(u_i)$ to whom user u_i replied only on topic T_Y.

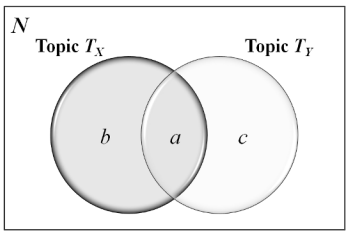

Figure 3: Observable quantities from user u_i's egocentric network $G'(u_i)$, considering the pair of topics T_X and T_Y.

It follows that the NPMI measure can be defined as:

$$NPMI(U_X, U_Y) = \frac{ln(\frac{a}{(a+b).(a+c)})}{-ln(a)}.$$

4. EXPERIMENTAL SETUP

The dataset used in this research was provided by sapo.pt[4] (the largest Web portal and ISP in Portugal) and comprises messages in Portuguese posted by Twitter users, between March 17^{th},2010 and May 16^{th},2010. Table 1 summarizes the volume of the dataset.

Table 1: Dataset description

Total number of users	49.303 users
Total number of tweets	612.556 tweets
Total number of replies	73.506 replies

Out of all the tweet messages, 12% are @replies, and these involve only 24% of the total number of users present in the dataset, which is 11.832 users. From the set of @replies we derived the *activity network*, representing the interaction among users.

Figure 4 depicts the distribution of the degree of the users in the *activity network*. The degree of user u_i corresponds to the number of connections u_i has in $G'(u_i)$. We observe that such distribution closely follows a power-law distribution. As expected, many users have low degree, meaning that they only communicate with a very small set of users, while few users have high degree (e.g., interactions with more than 10 users).

We also analyzed the frequency of the interactions among pairs of users, i.e., the weights of edges in the activity graph, finding that most users only interact occasionally. From Figure 5 we observe that the distribution of edge weights also follows a power-law. These observations are consistent with other works, even those involving larger-scale measurement in multiple social networks (see for example Mislove et al. [10]).

4.1 Topics Selection

To answer our research question, raised in section 3, we need to consider sets of user replies on different topics. Since the messages in our dataset were not categorized, we manually developed simple keyword-based classifiers to collect messages from three distinct topics: *Sports*, *Religion* and

[4]SAPO Labs: http://labs.sapo.pt/up/

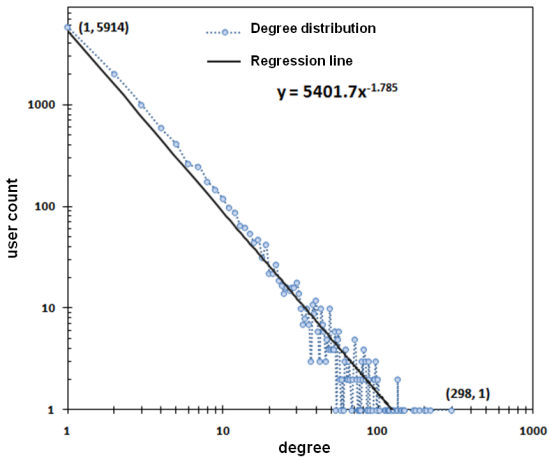

Figure 4: Distribution of node degree of the implicit network derived from tweet replies.

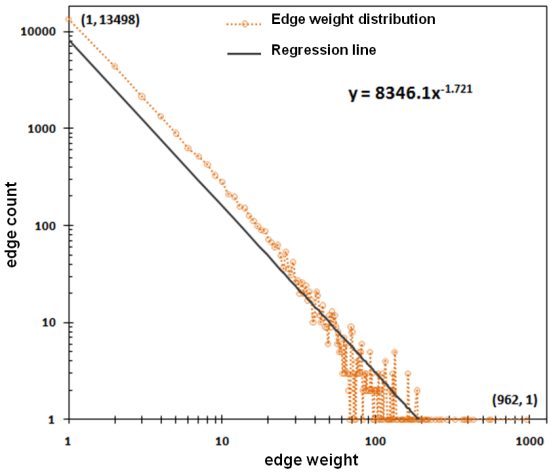

Figure 5: Distribution of edge weights of the implicit network derived from tweet replies.

Politics. Messages were considered to be about the Sports topic when they included keywords related with football teams, football players, football rules, managers and stadiums. Likewise, we used keywords like "God", "church", "faith", "saint", "sanctuary", "sin" and Popes' names to identify messages about Religion, and words such as "democracy", "monarchy", "crisis", "taxes", "laws" and names of politicians to identify messages about Politics. In total, we used a set of 58 keywords for the Sports topic, a set of 56 keywords for the Religion topic and 55 keywords for the Politics topic.

We selected these topics for several reasons. First, we did not expect these topics to be correlated. This is important, since we want to make sure that topics are sufficiently different to actually motivate users selection. Second, these topics were expected to be present in many tweets over time, i.e., we believed these topics would be recurrently discussed by the Twitter users. Figure 6 illustrates the distribution of messages in the Sports, Religion and Politics topics. These

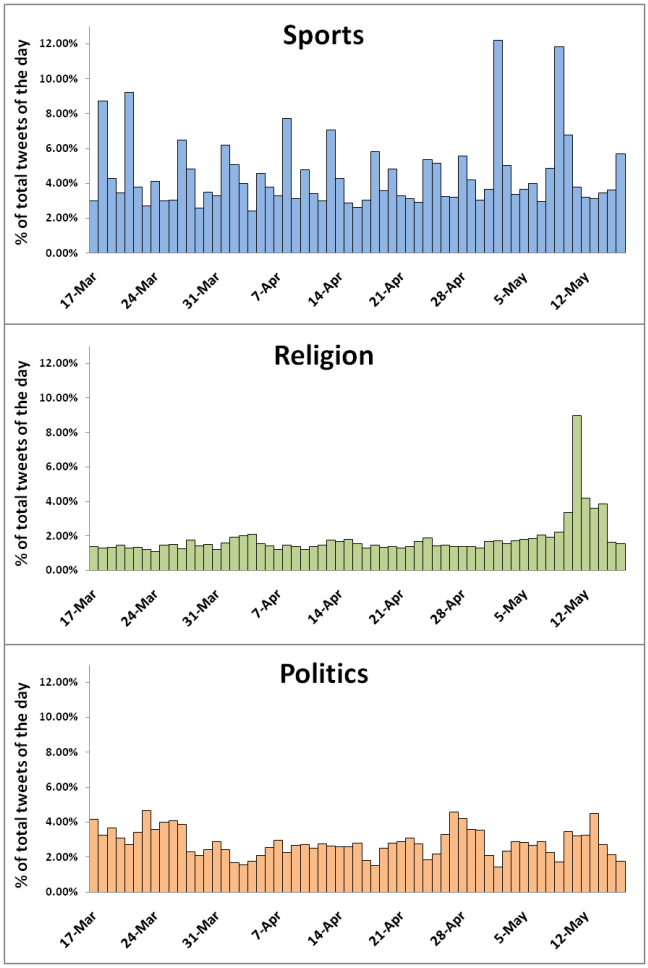

Figure 6: Distribution of tweets by day for *Sports* (top), *Religion* (middle) and *Politics* (bottom).

distributions confirm our intuition about the persistence of these three topics over time.

It should also be noted that the last two weeks – specially between May 3^{rd} and May 11^{th} – were marked by two relevant events that occurred in Portugal. One of them in Sports, occurred when the new champion of the national football league was found, while the other, in Religion, coincided with Pope Benedict XVI's visit to Portugal, a 4-day event between May 11^{st}-14^{th}, 2010. These events led to an increase in the number of posts in those specific days, as shown in Figure 6.

We also verified that the list of keywords used in the segmentation of the dataset by topic was sufficiently discriminating, since there are few replies on the intersections of the pairs of topics. As shown in Figure 7, the number of replies at the intersections of each pair of topics is significantly lower than the total number of messages in each topic. These values ensure that a user has a low probability of being included on the set of people that replied on a particular topic, without expressing a real interest in that topic. However, it is inevitable for some replies to fall into more than one topic. In fact, there are 3 tweets covering all three topics. We inspected these messages manually to verify that the

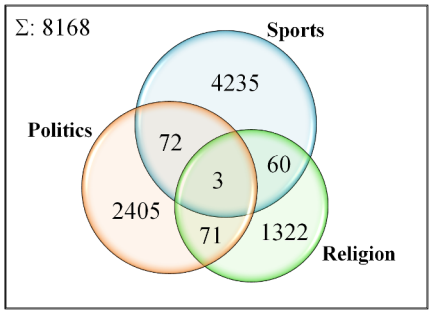

Figure 7: Number of replies on each topic.

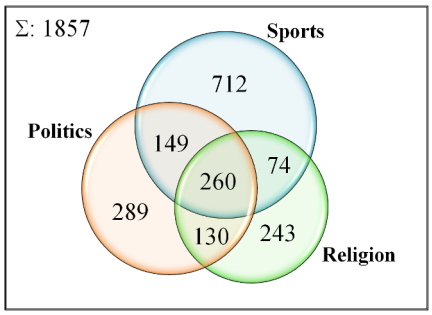

Figure 8: Number of users replying on each topic.

three topics were indeed present – the 3 tweets are provided in Table 2, showing the mixture of topics.

Focusing on the users, we observed there was a considerable number of users who replied to tweets in more than one topic, as shown in Figure 8. Thus, despite the fact that the topics are sufficiently orthogonal, many users interact in more than one topic. These observations ensure that our dataset is suitable for studying the users' motivations for replying to other people on these topics, through an analysis of the ego-centric networks in each topic.

Table 2: Tweets at the intersection of the three topics. The messages reference the football champion (i.e., "slb"/"benfica"), the Pope's visit to Portugal (i.e., "Pope") and also politics-related terms (i.e., "VAT tax", "government", "taxes" and "crisis").

Com **SLB Campeão**, as escolhas de treta do queiroz e SS o **Papa** em Portugal, o socas pode subir o **IVA** para 30%... etc. k ninguem nota!
Nao culpes o **slb** pelos **impostos**... Culpa o **governo** e a **crise**... E por falar em **slb** andavam uns malucos no marques a apitar... o **papa**!
E não te esquecas do **benfica campeão** e da visita do **papa**. Tudo para alegrar o povo. Acabou-se a **crise**!!

5. RESULTS AND ANALYSIS

In this section, we present our experiments and results with each pair of topics: Sports & Politics; Sports & Religion and Religion & Politics. For each pair of topics, we analyze the ego-centric networks of users who have replied

to messages from both topics, previously modeled as T_X and T_Y. There are users for which the corresponding topic sub-graphs – $G'_{T_X}(u_i)$ and $G'_{T_Y}(u_i)$ – do not overlap (excluding u_i), i.e., the sets of users U_X and U_Y are mutually exclusive. Thus, for each of the pairs of topics under evaluation, we first quantify the overlap. Then, for users with a non-empty overlap between the topic sub-graphs $G'_{T_X}(u_i)$ and $G'_{T_Y}(u_i)$, we compute the corresponding NPMI values.

5.1 Sports and Politics

Figure 9 shows the distribution of users with an ego-centric network of a given size, who replied on Sports & Politics (409 users), in two cases: i) the top distribution only includes users that totally segment their connections by topic (49% of the users), i.e., the cases where there is an empty overlap between the users in $G'_{T_X}(u_i)$ and $G'_{T_Y}(u_i)$; ii) the bottom distribution includes the remaining 51% of the users that replied to at least one other user on both topics. We observe that about half of the users with non-overlapping topic sub-graphs have essentially a small number of connections comprising only two or three users. On the other hand, users with larger ego-centric networks tend to have an overlap between the two sets of people for the current pair of topics.

The scatter plot presented in Figure 10 presents the values of the NMPI measure for each user with overlapping sub-graphs, i.e., case ii) above. It shows that, although there are few users who segment their social network based on the topic ($-1 \leq \text{NPMI} < 0$), the majority of users reply to messages about both topics from the same people, more than what was supposed to occur under the independence assumption ($0 < \text{NPMI} \leq 1$). The line plot represents the arithmetic mean of NPMI, and it shows the trend of the NPMI measure with respect to the size of the ego-centric networks. We also observe that the NPMI has a slight tendency to decline for users with less than 11 connections. It should be noted that the average NPMI values for ego-centric networks of size 11 or above correspond to particular cases, such as averaging over a small number of users (5 or less users). This leads the average to follow closely the distribution of NPMI. Thus, is not possible to generalize from this point on (i.e., size ≥ 11).

Overall, these results reveal that, for this pair of topics, users with larger ego-centric networks have a tendency to specialize their connections according to topics. However, the social aspect predominantly conditions users' interactions since the NPMI is almost always positive.

5.2 Sports and Religion

The Sports & Religion pair of topics involves messages from 334 users. Out of these users, 51% exhibit mutually exclusive topic sub-graphs, $G'_{T_X}(u_i)$ and $G'_{T_Y}(u_i)$, while the remaining 49% of users reply to at least one other user on both topics, leading to overlapping sub-graphs. Users with mutually exclusive sub-graphs are, in most cases, those with very small ego-centric networks, as depicted in Figure 11. With respect to the NPMI, the results resemble those obtained with the previous pair of topics. From the scatter plot of the NPMI and its average value for a given network size, shown in Figure 12, we observe that users with large ego-centric networks present a slight tendency to separate the network accordingly to the topic (i.e., NPMI tends to decrease). Again, we only consider this trend for ego-centric networks for which the average NPMI value has some significance (network size lower than 11).

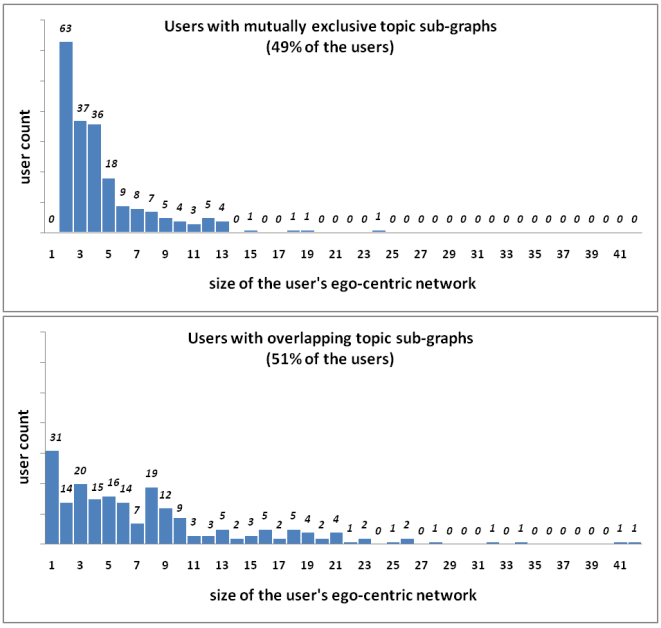

Figure 9: Distribution of users by ego-centric network size, for users replying on the pair of topics Sports and Politics, who have mutually exclusive (top) and overlapping topic sub-graphs (bottom).

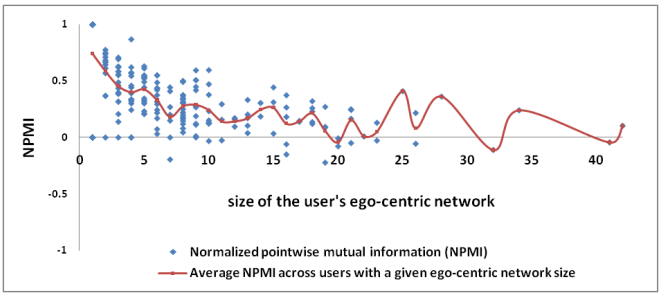

Figure 10: NPMI values for users replying in the pair of topics Sports and Politics.

5.3 Religion and Politics

In the Religion & Politics pair of topics (involving a total of 390 users), the split between users with mutually exclusive and overlapping topic sub-graphs is exactly 50%. As shown in Figure 13, users with mutually exclusive topic sub-graphs are again essentially those with a small number of connections. As observed with the other pairs of topics, the NPMI measure is in most cases positive. Figure 14 shows that the social aspect clearly prevails as the motivation for replying on these topics (i.e., $NPMI > 0$). Furthermore, it is clear that the users with larger ego-centric networks exhibit a stronger social character than the users involved in the previous pairs of topics. This could mean that the Religion & Politics topics are more likely to be of general interest and consequently,

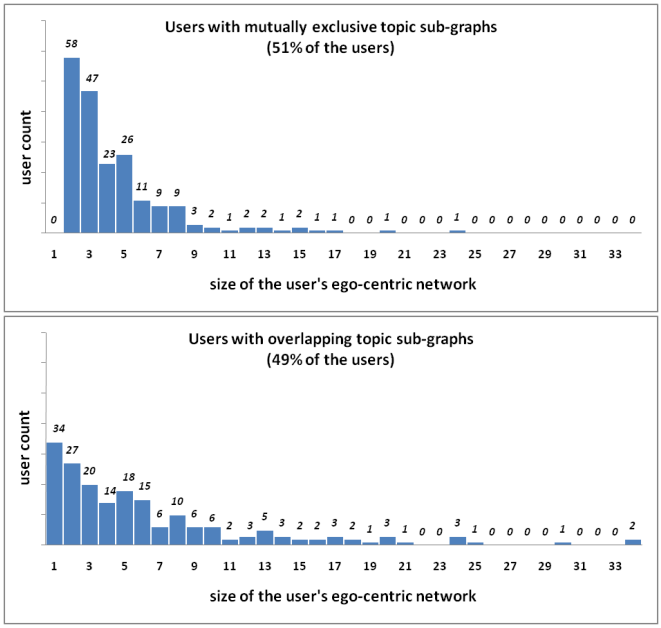

Figure 11: Distribution of users by ego-centric network size, for users replying on the pair of topics Sports and Religion, who have mutually exclusive (top) and overlapping topic sub-graphs (bottom).

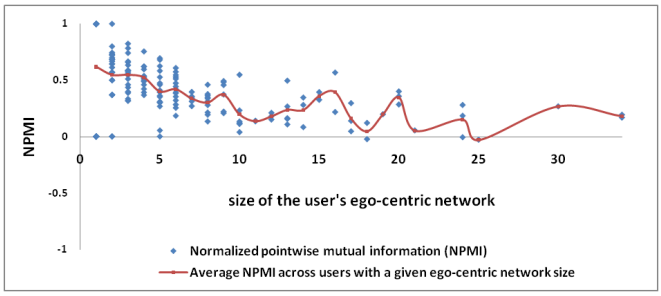

Figure 12: NPMI values for users replying in the pair of topics Sports and Religion.

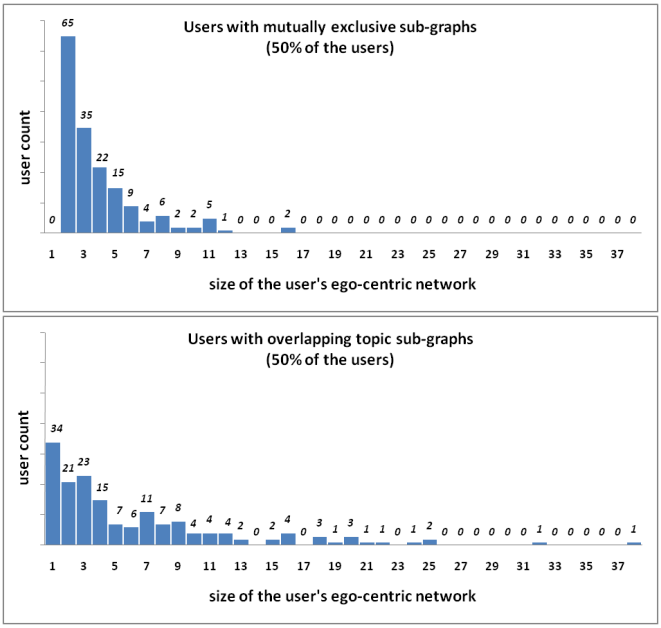

Figure 13: Distribution of users by ego-centric network size, for users replying on the pair of topics Religion and Politics, who have mutually exclusive (top) and overlapping topic sub-graphs (bottom).

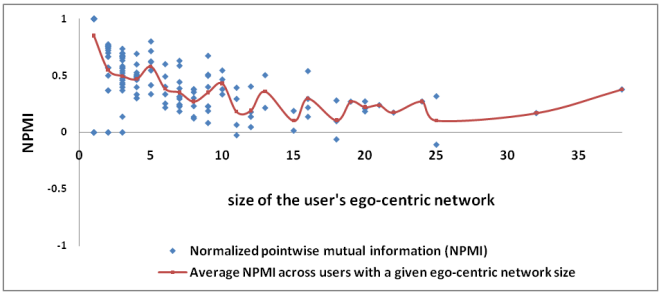

Figure 14: NPMI values for users replying in the pair of topics Religion and Politics.

lead users to be less topic-specific in their interactions. This might also be due to the fact that Sports seems to be, above all, the topic in which users are more topic-specific. Looking back at Figure 8, we see that the number of users that only reply on the Sports topic (i.e., 712 users) is substantially higher than the number of users that only reply in any of the other two topics (i.e., 289 users in Politics and 243 users in Religion). However, in general, the results obtained with this pair of topics follow the same trend to those obtained with Sports & Politics and Sports & Religion.

6. DISCUSSION

Our experiments with three pairs of topics reveal that users with smaller ego-centric networks are strongly influenced by the social aspect when replying to other individuals. The likelihood of such users discussing any matter with any of their connections is higher than for users with larger ego-centric networks. Moreover, as expected, when the number of connections that an individual has increases, there is a slight tendency to disjoint the network by topic. Still, the social aspect prevails over the topic as the motivation for interacting with other users, since the majority of users have a positive NPMI value. This result is not surprising and confirms our initial intuition.

Finally, we can say that, in addition to common measures of social network analysis that enable the characterization of individuals (e.g., degree, clustering coefficient, betweenness, closeness and eigenvector centrality), the NPMI measure could be used as a feature in a user classification process and could support processes of user profiling. The NPMI enables us to understand what motivates users to reply to other individuals: the topic or the social aspect. Thus, our research contributes with a particular methodology to analyze the topics' influence on users' interactions. Although

our results reveal that users with more connections have a tendency to separate the network by topic, we do not attempt to generalize these findings due to scale of our dataset and also to the "manual" process used for topics' selection.

7. CONCLUSIONS AND FUTURE WORK

In this paper we have presented a study of users' interactions through an analysis of the Twitter @replies network, using a dataset consisting of Twitter messages from Portugal. We focused exclusively on the *activity network*, i.e., the implicit social network derived from tweet replies, to investigate what motivates users to reply to other individuals. We showed that this implicit network presents some properties shared by many social networks. More importantly, we found that the social aspect appears to be predominant in motivating users' interactions. However, for users with larger ego-centric networks we observed a slight tendency for users to separate their connections depending on the topics discussed. Understanding the dynamics of networks and analyzing the topics' influence on user interactions are important to improve recommender systems, as we can customize the recommendations according to the user's profile.

As future work, we will be studying a larger set of topics and also comparing seasonal topics with permanent ones. Moreover, we will explore some popular techniques such as Support Vector Machines (SVM) [7] classifiers aiming to produce a classified collection, which will automate the process of network segmentation. Furthermore, instead of a manual topic selection, we will explore topic modeling techniques, such as Latent Dirichlet Allocation (LDA) [1] to discover the latent topics present in the dataset. This will allow us to replicate this study at a much larger scale.

8. ACKNOWLEDGMENTS

We would like to thank SAPO Labs (http://labs.sapo.pt) for providing access to the Twitter data crawl, and to Gustavo Laboreiro for the help in the preparation of the dataset used in the experiments. We would also like to thank the useful comments made by the anonymous reviewers.

9. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] D. Boyd, S. Golder, and G. Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *HICSS-43*, 2010.

[3] B. Gerlof. Normalized (pointwise) mutual information in collocation extraction. In *Chiarcos Eckart de Castilho & Stede, From Form to Meaning: Processing Texts Automatically*, page 3140. Proc. of the Biennial GSCL Conference, 2009.

[4] E. Gilbert and K. Karahalios. Predicting tie strength with social media. In *Proc. of the 27th International Conference on Human Factors in Computing Systems*, pages 211–220, Boston, MA, USA, 2009.

[5] S. A. Golder, D. M. Wilkinson, and B. A. Huberman. Rhythms of social interaction: Messaging within a massive online network. *Proc. 3rd Intl. Conf. on Communities and Technologies*, 2007.

[6] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proc. of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, pages 56–65, San Jose, CA, 2007.

[7] T. Joachims. Text categorization with support vector machines: learning with many relevant features. In *Proceedings of ECML-98, 10th European Conference on Machine Learning*, pages 137–142, Chemnitz, DE, 1998. Springer Verlag, Heidelberg, DE.

[8] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 611–617, Philadelphia, PA, USA, 2006.

[9] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *WWW '10: Proc. of the 19th International Conference on World Wide Web*, pages 591–600, New York, NY, USA, 2010.

[10] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *IMC '07: Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, pages 29–42, San Diego, CA, USA, 2007.

[11] M. Moricz, Y. Dosbayev, and M. Berlyant. PYMK: friend recommendation at myspace. In *Proc. of the 2010 International Conference on Management of Data*, pages 999–1002, Indianapolis, Indiana, USA, 2010.

[12] O. Phelan, K. McCarthy, and B. Smyth. Using twitter to recommend real-time topical news. In *Proc. of the 3rd ACM Conference on Recommender Systems*, pages 385–388, New York, NY, USA, 2009.

[13] J. Scott. *Social Networks Analysis: a handbook*. SAGE Publications Ltd, 2^{nd} edition, 2000.

[14] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi. On the evolution of user interaction in facebook. In *Proc. of the 2nd ACM Workshop on Online Social Networks*, pages 37–42, Barcelona, Spain, 2009.

[15] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1^{st} edition, Nov. 1994.

[16] J. Weng, E. Lim, J. Jiang, and Q. He. TwitterRank: finding topic-sensitive influential twitterers. In *Proc. of the 3rd ACM International Conference on Web Search and Data Mining*, pages 261–270, New York, NY, USA, 2010.

[17] D. Zhao and M. B. Rosson. How and why people twitter: the role that micro-blogging plays in informal communication at work. In *Proc. of the ACM 2009 International Conference on Supporting Group Work*, pages 243–252, Sanibel Island, FL, USA, 2009.