CHARACTERIZING ONLINE SOCIAL MEDIA: TOPIC INFERENCE AND INFORMATION PROPAGATION

by

SAED REZAYIDEMNE

A THESIS

Presented to the Department of Computer and Information Science
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Master of Science

March 2018

THESIS APPROVAL PAGE

Student: Saed Rezayidemne

Title: Characterizing Online Social Media: Topic Inference and Information Propagation

This thesis has been accepted and approved in partial fulfillment of the requirements for the Master of Science degree in the Department of Computer and Information Science by:

Reza Rejaie	Chair

and

Sara D. Hodges	Interim Vice Provost and Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded March 2018

THESIS ABSTRACT

Saed Rezayidemne

Master of Science

Department of Computer and Information Science

March 2018

Title: Characterizing Online Social Media: Topic Inference and Information Propagation

Word-of-mouth communication is a well studied phenomenon in the literature and content propagation in OSNs is one of the forms of WOM mechanism that have been prevalent in recent years specially with the widespread surge of online communities and online social networks. The goal of this study is to investigate what factors contribute into the propagation of messages in Google+. To answer to this question a multidimensional study will be conducted. On one hand this question could be viewed as a natural language processing problem where topic or sentiment of posts cause message dissemination. On the other hand the propagation can be effect of graph properties i.e., popularity of message originators or activities of communities. Other aspects of this problem are time, external contents, and external events. All of these factors are studied carefully to find the most highly correlated attribute(s) in the propagation of posts.

CURRICULUM VITAE

NAME OF AUTHOR: Saed Rezayidemne

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR, USA

Amir Kabir University of Technology, Tehran, IR

Mazandaran University, Babol, IR

DEGREES AWARDED:

Master of Science, Computer Science, 2018, University of Oregon

Master of Science, Electrical Engineering, 2011, Amir Kabir University of Technolog

Bachelor of Science, Electrical Engineering, 2008, Mazandaran University

AREAS OF SPECIAL INTEREST:

Online Social Networks
Content Propagation
Applied Machine Learning
Behavioral Analysis
Assistive Technology

PROFESSIONAL EXPERIENCE:

Researcher, Iran Telecommunication Research Center, 2011-2012

Network Administrator, TAM Iran Khodro, Tehran, Iran, 2012-2013

Graduate Teaching Fellow, Department of Computer Science, University of Oregon, 2013-2014

Graduate Research Fellow, Department of Computer Science, University of Oregon, 2014-2017

Researcher and Developer, Bionic Sciences Inc., Atlanta, Georgia, 2017-2018

PUBLICATIONS:

S. Rezayi, N. V. Parrish, S. Mirbozorgi, and M. Ghovanloo, "On Dual Band Considerations for Health and Technology Assistive Devices," *40th International Conference of the IEEE Engineering in Medicine and Biology Society*, to be submitted, July 2018

S. Rezayi, M. Sotoodeh, and H. Esmaili, "A Password-Based authentication and Key Agreement Protocol for Wireless LAN Based on Elliptic Curve and Digital Signature," *International Journal of Computer Science and Information Security* 9.10 (2011)

S. Rezayi, A. Hosseini, and H. Teheri, "Implementation of Extensible Authentication Protocol in OPNET Modeller," *Network, Communication and Computing (ICNCC), International Conference on. ACM*, 2011.

A. Hosseini, **S. Rezayi**, and H. Taheri, "Improving ertPS Grant Allocation for VoIP Traffic inSilence Duration," *International Journal of Information and Electronics Engineering* (2012)

A. Madani, **S. Rezayi**, and H. Gharaee, "Log management comprehensive architecture in Security Operation Center (SOC)," *Computational Aspects of Social Networks (CASoN), International Conference On. IEEE*, 2011.

N. Taherinejad, S. A. Mirbozorgi, and **S. Rezayi**, "Tuning the Harmony Memory Considering Rate (HMCR) parameter of Harmony Search Algorithm," *In 2nd Joint Conference on Intelligent Systems and Fuzzy Systems*, 2008.

R. Motamedi, **S. Rezayi**, R. Rejaie, R. Light, and W. Willinger, ""Who's Who" in Twitter: A First Look at the Twitter Elite Network," **to be submitted** to *ICWSM* Jan 2018.

S. Rezayi, R. Rejaie, "Characterizing Information and User Behaviour in Online Social Networks Through Text Mining." *University of Oregon Grad Forum poster presentation*, 2014.

S. Rezayi, R. Rejaie, "Information Propagation in Google+," *University of Oregon Grad Forum poster presentation*, 2015.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Reza Rejaie, without whom this work would not be possible. I would also like to thank Bahador Yeganeh and Soheil Jamshidi in the ONRG Lab, and Finally, I would like to thank my wife and my family for their support and inspiration.

To Mona

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

Figure Page

LIST OF TABLES

Table Page

CHAPTER I

INTRODUCTION

Growing levels of interactions between individuals and organizations through online social networks such as Twitter or Facebook has turned them into online information societies where users generate, propagate, exchange, receive information and act on it. Thus, there is a growing interest in mining this information for various purposes such as marketing, health, security, economics, etc. Paul and Dredze (2011), Bonchi, Castillo, Gionis and Jaimes (2011), and Tumasjan, Sprenger, Sandner and Welpe (2010).

Another body, understanding how users connects and interact on these systems is of great interest. Most studies have focused on the characterization of friendship structure among users. However, this does not offer any insight about the level of activity between users such as number and type of exchanged messages between users. Most exchanged messages are casual (commenting on a picture) that may not have a significant social or cultural implications.

In this work, we first show how to mine and characterize textual data from Online Social Networks and then we explore how this content spreads over the network.

CHAPTER II

TEXT MINING IN SOCIAL MEDIA

Introduction

Extracting information from online sources is challenging because length of a post is often short (for tweets it is 140 characters), and a post could be inherently ambiguous. Besides, use of unconventional language and unclear words and abbreviations adds to the complexity of analysis. One basic issue for information mining is to provide some basic context for a post, such as its topic. More specifically, given a post, can we infer whether it is about soccer, politics, etc. However, There is no widely accepted set of a topics with a clear granularity (e.g. what is a proper granularity for a topic, should we consider sport or soccer as a topic). This issue and the fact that posts could be too simple (no topic) or too complicated (multiple topics) makes the problem more challenging.

Machine learning techniques are promising approaches for such inferences. Prior studies have used Topic Modeling to find a topic of a document. However these algorithms are highly dependent on the number of topics. It might be impossible to figure out the right number of latent topics in LDA Algorithm Blei, Ng and Jordan (2003) and such number may not even exist. We address this issue in Section II. As a result, our goal is to infer a topic of a post using supervised classification.

However, before pursuing our goal we would like to investigate topics of tweets as they are perceived by humans. To make this manageable consider a case with N specific topics of interest. Toward this end we use categories used

by an online marketing website namely socialbakers.com and we collect tweets of well known accounts per category. To tackle the challenge in supervised learning we label the tweets by humans and our hypothesis is that "professional accounts generate tweets related to their category." For example consider the following tweet: "LIVE: President Obama is speaking at the White House" put out by the account Barack Obama. We can intuitively say that Barack Obama falls into the politics category and also its tweet has the topic of politics. That is why we first study whether individual tweets have clear and unique topics as they are perceived by humans rather than simply using a supervised LM technique.

We would like to gain insight about following fundamental questions:

- *How are topics of tweets perceived by humans? Do tweets have one or multiple or no clear topics?* The answer to these questions is important because a tweet is our only source of information that we use to train our model and if we do not train the system precisely how could we except that machine assigns a topic to a short text that has no information in it, "Enjoy the sunshine" for instance!

- *To what extent is topic of a tweet aligned with the category of the account that generated the tweet?* The answer to this question could vary across different categories and even among accounts in a single category. In fact, the alignment of tweet topics with category of an account shows how that entity associated with the account is using Twitter, e.g. announcement, advertisement, voting media, etc.

– *How do professional Twitter accounts use Twitter?* As a result of the above question we are also interested in answering this question.

The rest of this chapter is organized as follows: Section II reviews the related works in this area. Section II presents data collection and data labeling and a summary of our dataset. Sections II characterizes the dataset and investigates the alignment of account category and tweet topic. Also we present our feature set that is used for rule based classification in this section. Section II then leverages classification technique to infer a topic from a tweet. Section II investigates if there are certain keywords that are related to different categories. Finally, Section II presents our conclusions.

Related Work

Assigning a topic to a document is not a new problem and there have been many efforts in analyzing social network text. In general, there are two approaches for natural text processing: unsupervised and supervised analysis. Unsupervised analysis is generally called clustering that divides a set of objects into clusters so that objects in the same cluster are similar to each other. These algorithms, e.g. K-means Hartigan and Wong (1979), are unsupervised, meaning no humans input is necessary. Topic inference has plenty of application from recommender systemsWang and Blei (2011) to ad placement Ahmed, Low, Aly, Josifovski and Smola (2011) and interest miningGuy et al. (2013).

All studies in this domain are categorized under Machine Learning (ML) techniques. To analyze text and retrieve information from it, classification have been widely used and studied where a model is trained by a set of pre-labeled

documents (training set) and is asked to classify a new set of unseen documents (test set). Koller and Sahami (1997), Joachims (1998), and Yao, Mimno and McCallum (2009), have leveraged popular classifiers on text.

There are other studies that use classification to infer other properties of tweets like sentiment analysis in Gonçalves, Araújo, Benevenuto and Cha (2013) and Kouloumpis, Wilson and Moore (2011) or measuring question quality in Zhao and Mei (2013) or link prediction Barbieri, Bonchi and Manco (2014); however the limited information in Twitter text (each tweet is limited to 140 characters) has caused difficulties in the task of topic inference.

There is another emerging technique called topic modeling that can be supervised Blei and McAuliffe (2007) or unsupervised Purver, Griffiths, Körding and Tenenbaum (2006). These algorithms discover semantic structure of documents, by examining word statistical co-occurrence patterns within a corpus of training documents. Authors in Hong and Davison (2010) address the problem of using standard topic models in micro-blogging environments (such as Twitter) by studying how the models can be trained on the dataset. L-LDA (Labelled LDA) that is proposed in Ramage, Hall, Nallapati and Manning (2009) is based on LDA Blei et al. (2003) and is a supervised topic model for assigning topics to a collection of documents.

Data Collection and Data Labeling

This section describes our dataset and the way we label tweets. All general statistics are provided here including number of categories, number of accounts per category and number of labeled and unlabeled tweets per account.

Tweet Collection. To build an effective training set, we select a group of Twitter accounts that are related to a specific *category* [1] and collect all available tweets from these accounts. This approach to data collection not only increases the likelihood of collecting tweets that are related to the selected categories but also enables us to examine to what extent the topic of generated tweets by individual accounts are related to the category of the account. Toward this end, we use web sites, namely socialbakers.com, that publish list of popular Twitter accounts (including their Twitter IDs and the number of followers) that are classified into more than 80 categories. We identify 16 categories and hand pick a set of accounts that represent well known entities (i.e.major teams, companies, brands with a large number of followers) for that category.

While focusing on well-recognized accounts may limit the number of selected accounts in some categories, it intuitively increases the likelihood that their tweets are related to their category as their accounts are likely to be professionally managed. The selected categories essentially define the scope of our study. The list of selected categories along with the number of related accounts and collected tweets in each category is summarized in Table 1. The complete list of all selected accounts for each category and their associated tweets is available in the Appendix 2.

While our goal is to ensure that selected categories are clearly separated, achieving this goal is not trivial. Intuitively, there is some overlap between pairs of selected accounts (e.g.fashion and beauty, or beverage and alcohol), and a category

[1]Throughout this paper, we use the term "category" to refer to the context of individual Twitter account, and use the term "topic" to indicate the context of individual "tweets". Using different terms should further clarify the focus of each discussion.

topic	No of accounts	No of tweets	No of tweets with one label	No of tweets with three labels
airline	10	32,229	5,393 (%16.7)	600 (%1.8)
alcohol	10	28,339	5,398 (%19.0)	599 (%2.1)
auto	12	38,589	6,472 (%16.7)	720 (%1.8)
basket	9	28,850	4,848 (%16.8)	540 (%1.8)
beauty	10	32,211	5,362 (%17.6)	596 (%1.8)
beverage	10	32,969	5,362 (%16.2)	599 (%1.8)
education	11	33,773	5,923 (%17.5)	655 (%1.9)
electronics	12	37,522	6,494 (%17.3)	720 (%1.9)
fashion	14	34,837	7,109 (%20.0)	702 (%2.0)
finance	11	31,776	5,391 (%16.9)	598 (%1.8)
gaming	6	19,383	3,209 (%16.5)	357 (%1.8)
health	10	27,726	5,395 (%19.4)	599 (%2.1)
news	14	45,044	7,575 (%16.8)	840 (%1.8)
politics	15	36,923	7,722 (%20.9)	781 (%2.1)
soccer	12	38,522	6,175 (%16.0)	677 (%1.7)
telecom	7	22,583	3,775 (%16.7)	420 (%1.8)
total	173	521,276	91,603 (%17.5)	10,003 (%1.9)

Table 1. List of selected topics and fraction of single/multiple-label tweets

such as news has inherent overlap with a few other categories (politics, finance, or sport). Considering these overlapping categories enables us to explore the potential effect of category overlap on our analysis.

Tweet Labeling. We recruited a group of UO students to specify the topic (i.e.label) of a subset of tweets in our dataset. Toward this end, each student is provided with a spreadsheet that includes the text of a random selection of tweets and prompts them to assign a topic to each tweet from a drop-down menu. This menu of topics contains all sixteen categories along with two more sensible categories: "no topic" and "other". Students are instructed to assign the label "other" to a tweet if it has a pronounced topic that is not listed in the menu (e.g.music), and assign the label "no topic" if they can not associate any clear topic to a tweet (e.g."2010 has been an exception year").

The assigned tweets to students are organized into two mutually exclusive groups:

- Three label tweets: Tweets that were labeled by three different students
- Single label tweets: Tweets that were labeled only once.

The multi-label tweets enable us to examine the consistency of label assignment by individuals. Such an inconsistency could be due to genuine disagreement among students on the topic of the tweet or caused by mistakes. The last two columns of Table 1 specifies the fraction of tweets (for each category) that has been labeled once or three times. As this table shows, the recruited students have assigned more than 121.6K labels (including 3 separate labels for 10K tweets).

For each tweet with three labels, we define the notion of Level of Agreement (LoA) that shows the maximum number of similar labels. More specifically, we use the term LoA3, LoA2 and LoA1 for a tweet with three labels to indicate that its number of similar labels are 3, 2, or 1, respectively. We also use the notation of LoA2+ to refer to the collection of tweets that have LoA2 or LoA3 (i.e.LoA2+ = LoA2 $\cup$ LoA3).

Characterizing Assigned Topics by Human Labels

We leverage the tweets with three labels to examine the characteristics of assigned topics to tweets by humans. These characteristics provide the basic understanding of the clarity of topic for individual tweets and the alignment between the topic of tweets and the category of their associated account. The obtained insights from these characterization effort will inform the evaluation of classification techniques in the second half of the paper.

The task of assigning a label to a tweet may not be trivial when the associated keywords offer diverse clues. For example, a tweet with keywords "Clare Choir, tour, Australia" provides clue about traveling, music and singing, as well as education (since Clare is a college at Oxford University). However, a person who does not know about the educational context, will not assign the label of education to this tweet. In essence, the available information and context to individuals could affect the way they perceive and thus label tweets with diverse clues.

Despite this challenge, having three labels for each tweet enables us to determine the topic of a tweet with relatively high confidence. In particular, we assume that if at least two assigned labels for a tweet are similar (i.e.any LoA2+

tweet), the common label determines the topic of the tweet since it is unlikely that two individuals make a similar mistake in assigning a label. Note that the common label of a tweet might be *aligned* or *misaligned* with the category of the corresponding account. For example a tweet that has these keywords "Reuters, US Econ, collapse, benefits, $29B, GM" which are associated with a Twitter account with the category of auto and has three similar labels of finance is a LoA3/misaligned.

Hence for each tweet we measure LoAi/x metric where i shows the level of agreement between labels ($i \in \{1, 2, 3\}$) and x indicates the alignment ($x \in \{aligned, misaligned\}$)

We have manually inspected hundreds of LoA2+ tweets to verify the use of common labels as the topic of tweets for LoA2+ tweets that are both aligned and misaligned with their corresponding accounts' category. We observed that for an absolute majority of LoA2+ tweets ($> 95\%$) the common label is the most reasonable topic. The most common exceptions are tweets whose common misaligned label is "no topic" or "other" due to the lack of a dominant context for the tweet. For example, a tweet with keywords "disaster, texting, Redcross" is associated with an account of health category but was labeled twice as "other". Our inspections confirm that the common label for LoA2+ tweets can reliably be used as the topic of the tweet despite stated challenge for humans to assign a consistent topic to tweets with conflicting clues. In the rest of this section, we characterize the topic of LoA2+ tweets in order to answer the following key questions:

– Does (and to what extent) the topic of the generated tweets by (professional) Twitter accounts is aligned with their category across different categories?

– Does the level of alignment between the category of a Twitter account and the topic of its tweets vary across different categories?

– What does the alignment between the category of an account and its tweets reveal?

Figure 1. Agreement between tweet labels and account category for three label tweets per account

Alignment of Account Category and its Tweet Topic. To explore the relation between the category of an account and the topic of its tweets, we divide all tweets of each selected account into the following three groups:

– LoA2+/aligned

– LoA2+/misaligned

– LoA1

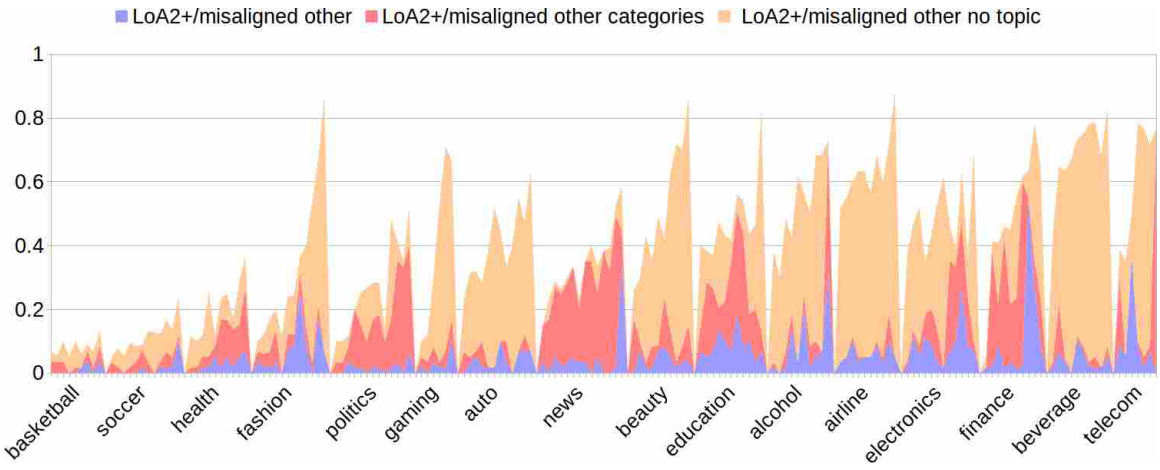


Figure 2. Breakdown of LoA2+/misaligned tweets among "other", "no topic", and "other categories" per account

We refer to these three groups as aligned, misaligned and ambiguous tweets, respectively. Intuitively, these three groups of tweets respectively indicate the extent that generated tweets by an account is related or unrelated to its category or is ambiguous. In essence, the specific division of tweets across these three groups can provide a valuable insight on how these Twitter accounts are used by their owners.

Figure 1 presents the percentage of tweets across these three groups for each account. Furthermore, accounts within the same category are bundled together, categories are ordered (from left to right) based on their average percentage of LoA2+/aligned and within each category accounts are ordered (from left to right) based on their percentage of LoA2+/aligned. This figure illustrates following interesting points:

First, there are some variations in the division of tweets among aligned, misaligned and ambiguous groups within each category. We observe that in some categories (soccer, basketball, health, politics) most accounts clearly exhibit a

much larger percentage of aligned tweets than other categories. We refer to these categories as *purposeful* as a significant fraction of their tweets are related to their mission. In contrast, in some other categories (telecom, beverage, finance, electronics, airlines, alcohol, education) a significant percentage of published tweets are misaligned. We refer to these categories as *aimless*. In essence, the relative percentage of aligned and misaligned tweets appears to be largely related to the category of the accounts. Second, the percentage of ambiguous tweets is around 10% to 30% in most cases and is relatively stable across different categories.

Misaligned Tweets. To gain more insight into the LoA2+/misaligned tweets, we take a closer look at this group by dividing them into the following three subgroups based on their inferred topic (that is misaligned with its category):

- *Other*: tweets whose label is "other"
- *No Topic*: tweets whose label is "no topic"
- *Other Topics*: tweets whose label is the same as one of the other 15 categories.

Note that the characterization of these misaligned tweets are more relevant to aimless categories as most of their tweets are misaligned.

Figure 2 plots the percentage of all LoA2+/misaligned tweets among the above three types for each account, i.e.essentially providing the breakdown of the LoA2+/misaligned in Figure1. This figure clearly illustrates that a significant fraction of misaligned tweets in some "aimless" categories, namely telecom,

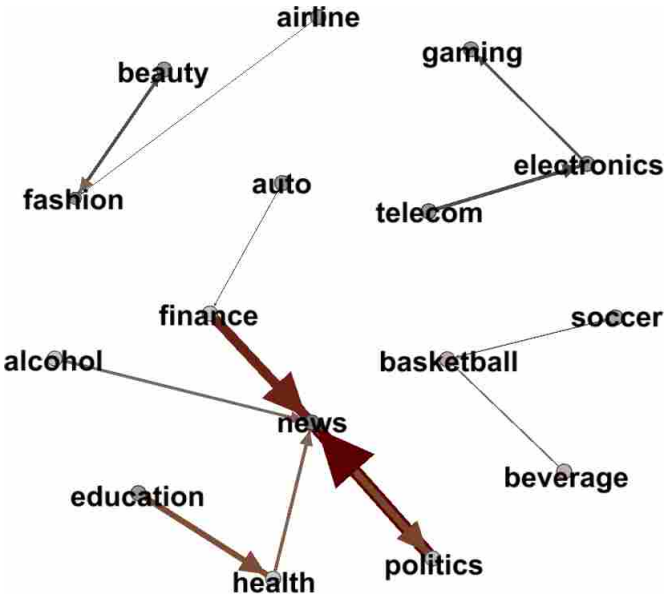


Figure 3. Other major related categories for multi purpose accounts.

beverage, airline, alcohol, beauty, auto and gaming, have no topic at all. This reconfirms our earlier assertion that these categories generally appear to be aimless.

In contrast, a majority of misaligned tweets in some other categories, namely finance, education, news, politics, and health are mapped to one of our other categories. We refer to these categories as *multi purpose* categories. In Figure 3 we try to visualize this metric as a graph. In this graph nodes are categories and edges are number of mislabeled tweets between to categories. As can be seen, edges are weighted and directed. Weight represents the number of mislabeled categories and is proportional to thickness. Direction shows in which way we have mislabeling. For example a large number of finance tweets are labeled as news but for news politics is the second major category. Accordingly we draw a conclusion that the edges between two categories shows the overlap between those two categories. This figure also clearly illustrates that news is a multi purpose category and it mainly has overlap with politics and finance. Another pair category is basketball and soccer

because they fall into super category of sport. For some sample tweets that shows the multi purpose nature of tweets see Table 2.

tweet	label1	label2	label3	category
Pro-Obama nonprofit will no longer divert gifts to allied groups	politics	politics	news	news
Wall Street is sharply divided on 2015 outlook [CNBC Fed Survey]	finance	finance	news	news
Follow the fragrance trail of Jadore from Grasse	beauty	beauty	beauty	fashion
@PlayStation: 12GB PS3 system will be $199 in North America.	gaming	gaming	gaming	electronics
Spurs Connect: Free App for Spurs fans Now on Android	soccer	basketball	basketball	soccer

Table 2. Sample tweets for LoA2+/misaligned with other categories that shows multi purpose nature of some categories.

Ambiguous Tweets. We now turn our attention to the LoA1 subset of tweets that have very diverse labels. To learn more about these tweets, we divide them into two more groups:

- *LoA1/aligned*: the tweets for which one of their labels is aligned with their category.
- *LoA1/misaligned*: the tweets that none of their labels is aligned with their category.

Figure 4 depicts the break down of the total percentage of LoA1 tweets for each account into LoA1/aligned and misaligned.

We can clearly observe that for many categories, an absolute majority of LoA1 tweets are LoA1/aligned with their category. This implies that tweet's

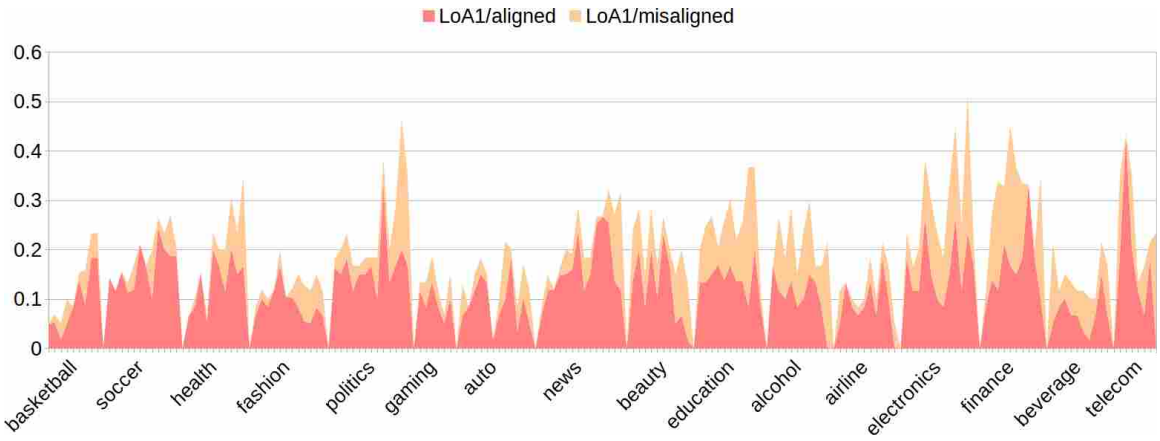


Figure 4. Breakdown of LoA1 tweets for each account into aligned and misaligned

context has some connection with its category but it may not very obvious/strong. Our closer inspection of these tweets revealed that most of these tweets can indeed be reasonably associated with two different topics, the third label is in some cases a very reasonable one and in other cases appear to be a mistake. To demonstrate this point consider the following LoA1/aligned tweets: *"Tories, Labour and Lib Dems to declare opposition to a currency union with Scotland"* with the account category of news that received three reasonable labels of news, politics and finance, or *"Download the new Fox News app for Android. Watch Fox News Channel live"* that has the category of electronics and was properly labeled as telecom, news, and electronics. However, this tweet *"Monica Lewinsky speaks out, says she was made scapegoat"* received two appropriate labels of politics, news and one seemingly in appropriate label of fashion while its category is news.

Automated Classification of Accounts. So far we have broadly classified Twitter accounts based on their LoAi/x characteristics in a hand crafted manner. Each account has a few LoAi/x numbers that can be viewed as its *features*. We can use a classifier to identify the rules for accounts in each category.

Obviously, the rules may not be perfect and some accounts are grouped with other categories. We use decision tree classifier to generate these rules and examine whether they are aligned with our earlier hand crafted classifications. This exercise also shows the relative distance between categories.

The list of features that are fed into decision tree classifier are as follows:

feature name	abbreviation
LoA2+/aligned	LoA2+/a
LoA2+/misaligned with other	LoA2+/mo
LoA+/misaligned with no topic	LoA2+/mnt
LoA2+/misaligned with other topics	loA2+/mot
LoA1/aligned	LoA1/a
LoA1/misaligned	LoA1/m

Based on the generated tree, LoA2+/a has the highest information gain and becomes the root for the tree and it splits all accounts into two imbalanced subgroups. The tree is generated graphically and is available in Appendix 1. Here we list some sample rules that show these features lead us to the correct point. Also Figure 5 is a part of this tree that reveals the following rules.

(LoA2+/a $>$ 45.8%) $\wedge$ (LoA2+/mot $>$ 9.16%) $\wedge$ (LoA2+/mo $>$ 1.68%) $\Rightarrow$ 60% politics

(LoA2+/a $>$ 45.8%) $\wedge$ (LoA2+/mot $>$ 9.16%) $\wedge$ (LoA2+/mo $<=$ 1.68%) $\Rightarrow$ 60% news

These rules confirm our previous observation in Figures 1 and 2. For example in Figure 2, we observed that LoA2+/misaligned with "other" categories has a great share of all LoA2+/misaligned tweets for news and politics, and classification place them in a same branch.

In another branch we see that finance and news has the same number of accounts in one leaf. In other words we can extract following rule:

(LoA2+/a <= 45.8%) ∧ (LoA2+/mnt <= 34.1%) ∧ (LoA2+/mot > 6.7%) ∧ (LoA2+/mo <= 5.8%) ⇒ 30% news and 30% finance which is consistent with Figure 3 that shows news and finance have the closest distance after news and politics.

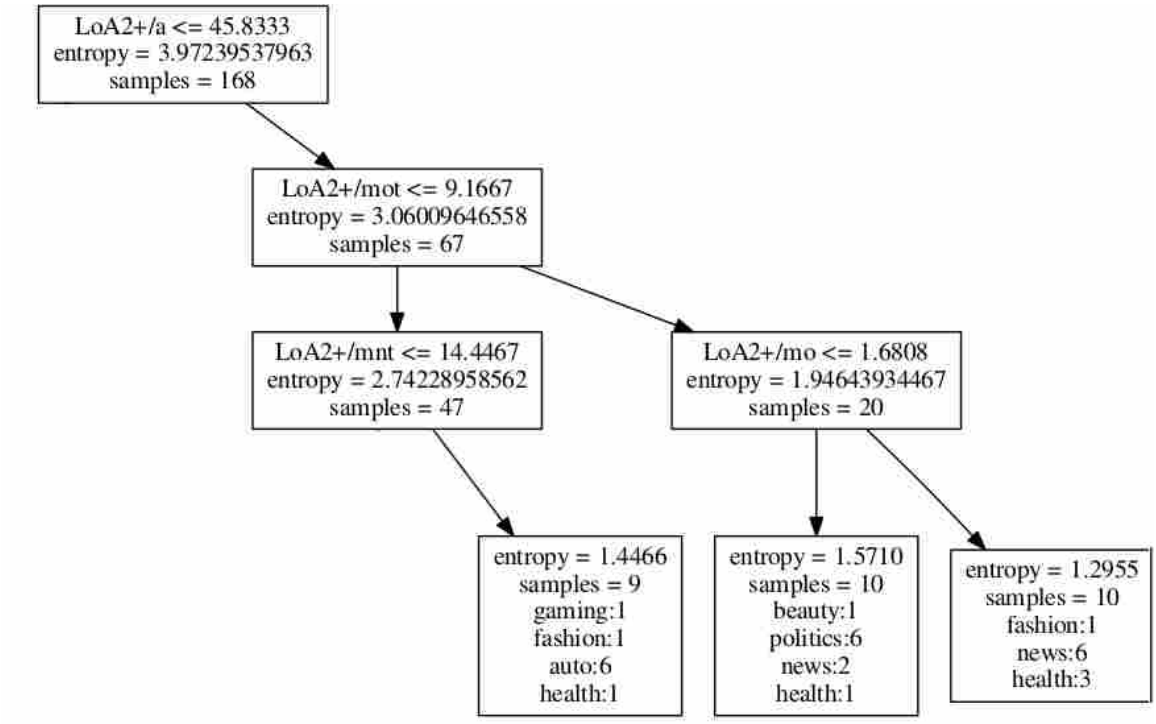


Figure 5. Partial decision tree for politics and news

Inferring Used Strategy by Accounts/Categories. As a result of above exercise we can elaborate on how certain accounts use twitter, (e.g.informing followers about deals, providing info, asking them to vote) and how this type of use is aligned with classification result (in Section II), and whether the accounts are managed professionally or casually.

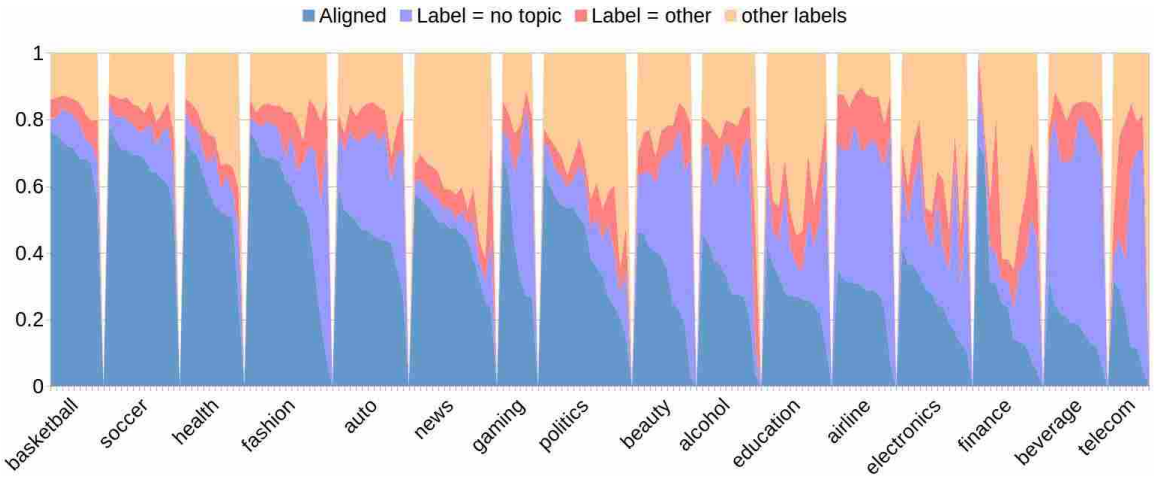


Figure 6. labeling information for single label tweets per account

According to the decision tree model, we see none of the leaves is clearly associated with category telecom as telecom accounts are scattered in four different leaves. This suggests that telecom accounts do not use Twitter for telecommunication reasons. We can verify this claim by manually checking the tweets of these accounts.

For example 65% of tweets of account Sprint is the following text!

Please visit *some url* to complete your contest entry!

where *some url* is a url that will be redirected to the sprint website when it is clicked.

Another telecom account Skype uses Twitter very casually and mostly to thank their costumers and ask about their feedbacks. We list some of its tweets in table **??**.

As it is seen nothing informative could be found in these tweets and we can not expect that machine or humans could infer an appropriate topic for this

Awesome! We're glad we can be there for you. :)
Wow, you must really love the emotions. Who do we help you stay in touch with? :)
glad we could bring a few extra laughs to your day. Do you and your brother catch up often?
We are here to help. :)
Sounds like someone was a little bit tired ;)
We're glad we can be a part of your daily ritual!

Table 3. Sample tweets for telecommunication account Skype

account. Such accounts can be found in other categories as well. Redbull is an example of beverage category that uses Twitter the exact same way as Skype does and no beverage related keyword could be found in its tweets.

In summary our characterization of labels reveals the clarity and complexity of topics of tweets as they are perceived by humans. We also examined alignment of tweet topic with category of each account. The insight of this section helps our automated topic inference in the next section.

Text-based Topic Inference of Tweets

We now turn our attention into the automated classification of tweets from the target account into one of the specified topics.

Dataset: To expand our dataset for this analysis, we use the larger set of single label tweets that are presented in Table 1. Figure 6 shows the division of tweets for each account across four groups based on their labels:

	case 1		case 2		case 3	
category	NB	SVM	NB	SVM	NB	SVM
soccer	**0.97**	0.95	0.75	**0.87**	**0.93**	0.92
airline	0.64	**0.87**	0.16	**0.71**	0.65	**0.68**
basketball	0.8	**0.84**	0.68	**0.77**	**0.7**	0.69
health	0.76	**0.83**	0.37	**0.68**	0.47	**0.60**
news	0.67	**0.76**	**0.88**	0.6	**0.75**	0.7
politics	**0.78**	0.77	0.28	**0.53**	**0.54**	0.53
fashion	**0.80**	0.7	0.47	**0.54**	**0.58**	0.46
beauty	0.21	**0.61**	0.04	**0.43**	0.42	**0.47**
gaming	0.13	**0.58**	0.05	**0.47**	**0.40**	0.38
auto	0.52	**0.58**	0.07	**0.47**	**0.47**	0.39
alcohol	0.26	**0.57**	0.07	**0.40**	0.41	**0.42**
education	0.1	**0.55**	0.01	**0.30**	**0.36**	0.34
electronics	0.1	**0.39**	0.02	**0.29**	**0.38**	0.28
finance	0.01	**0.31**	0.01	**0.24**	0.15	**0.16**
telecom	0	**0.23**	0	**0.21**	0.14	**0.16**
beverage	0.01	**0.17**	0	**0.19**	**0.34**	0.32

Table 4. Accuracy result for all classifiers and two datasets

- Aligned: tweets whose category and label agree.
- No topic: tweets that are labeled as "no topic".
- Other: tweets that are labeled as "other".
- Other labels: tweets that are labeled as one of the other categories.

Accounts of each category are grouped together. Categories are ordered from left to right based on their average percentage of aligned tweets and within each category accounts are ordered based on the same criteria. Therefore, Figure 6 is comparable to Figure 1. We observe that the order of categories and accounts in each category in Figure 1 and Figure 6 are exactly the same. Comparing these two figures reveals that three- and single-label tweets for each account exhibit generally similar characteristics.

Methodology. We only focus on English tweets and we use the *bag of words* approach to process these tweets. After filtering stop words, we consider all words of a tweet as features when feeding them to a classifier. Each word and similarly each tweet is assigned a unique ID. For each tweet, we count the number of occurrences of each word so we would have a $W \times D$ matrix where W is the number of distinct words and D is the number of documents (here each tweet is a document). For analyzing single label tweets whose label and category agree, the number of distinct vocabularies is 88,373 and the number of documents (tweets) is 36,559. Therefor, the size of the matrix is very large; however it is also very sparse (i.e. most values in matrix are zeros) and only non-zero values are stored. The only filtering that is implemented here is removing stop words.

Next, we use *tf-idf* – stands for *term frequency inverse document frequency* – weighting scheme Sparck Jones (1988) to produce a weight for each word. This weight is highest when the word w occurs many times within a small number of documents and vise versa. The *tf-idf* matrix then is fed to two well known classifiers in the area of text mining for building the model; *(i)* Support Vector Machine (SVM) and *(ii)* Naive Bayes (NB). Other classifiers such as Linear Regression, Ridge Classifier, and Nearest Centroid are also implemented, but since their results are not better than SVM we just report their accuracy here and do not go into their details. In the next subsection we cover briefly why we focus on these classifiers.

All classifiers are implemented in Python using SciKit library Pedregosa et al. (2011). We run the classifier on three different cases as follows:

Case 1: considering single label tweets whose label and category agree.

Case 2: considering all single label tweets leveraging only labels and ignoring categories.

Case 3: considering all tweets.

Note that the quality and reliability of specified topics for tweets decreases from Case 1 to Case 3. This allows us to study the effect of training set on classification accuracy which will be discussed in Section II.

In all these cases, we employ *leave-one-out* cross validation in which we use tweets of 172 accounts for training and the tweets of the remaining one account for testing. Therefor, we repeat this process 173 rounds for each case.

The main motivation for leave-one-out testing (instead of using random tweets) is to assess whether training a classifier by $n-1$ accounts per category leads to a good classification of tweets on the single test account. This shows whether the selection of testing accounts have impact on the classification accuracy.

Classifiers. Classification and regression are supervised learning techniques to create models for prediction. Regression is when we predict quantitative outputs, and classification is when we predict qualitative outputs Hastie, Tibshirani and Friedman (2001). By using a threshold, regression turns into classification, so in this text we use the terms classification and regression interchangeably.

Classifiers are grouped into two categories: Generative and Discriminative. A generative model is a full probabilistic model of all variables, whereas a

discriminative model provides a model only for the target variable(s) conditional on the observed variables.

Generative Classifiers: The way generative classifiers work is to model how the data is generated. Then based on generation assumptions, find the class which is most likely to generate the test data. These classifiers explicitly model the actual distribution of each class. One popular classifier in this category is Naive Bayes. This classifier applies Bayes Theorem to distinct between different classes. For the text data, usually word count is considered as a feature, and it is called *naive* because it assumes that the value of a particular feature is unrelated to the presence or absence of any other features.

Discriminative Classifiers: Discriminative algorithms allow to classify points without providing a model of how the points are actually generated. In short, discriminative classifiers try to model the decision boundary between the classes. Support Vector Machine is a typical discriminative classifier. It constructs a set of hyperplanes in space and tries to find a separator between samples, That are called support vectors. SVM does not try to understand the basic information of the individual classes as Naive Bayes does. Ridge Classifier, Nearest Centroid, and Linear Regression are other popular discriminative classifiers that have shown an acceptable performance in text data, which is why we implement them here in this project.

A. Jordan in Jordan (2002), which is a widely cited study on the subject of discriminative vs. generative classifiers, compares Naive Bayes with Linear Regression. This study shows that discriminative models generally outperform

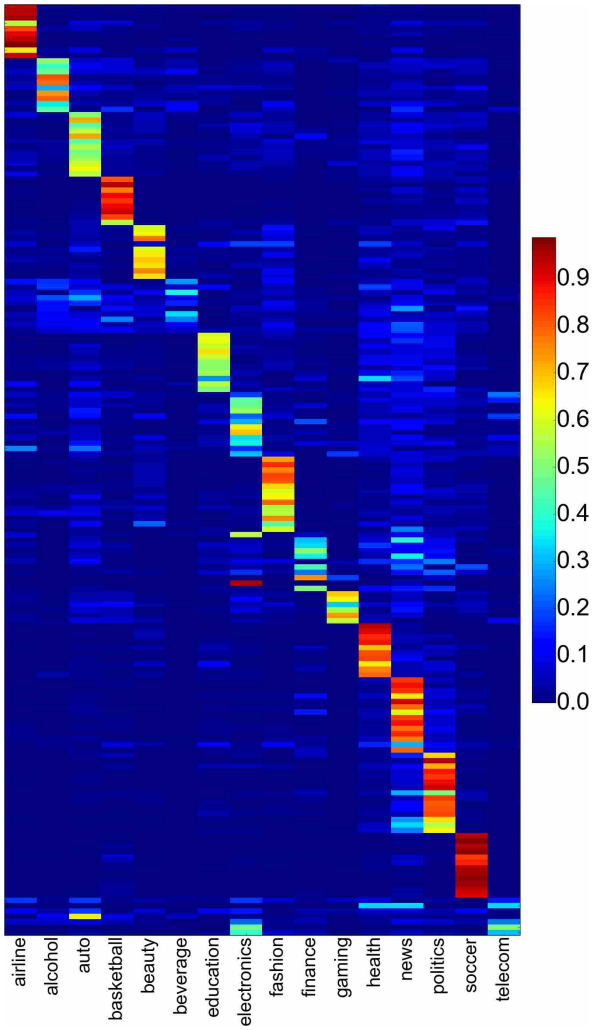


Figure 7. Account based accuracy heat map for support vector machine case 1

generative models in classification tasks in terms of accuracy but fall behind from generative classifiers in terms of convergence rate.

Per Category Analysis. We first examine the accuracy of classifiers at the per category level. Using leave-one-out cross validation, we measure the accuracy of each classifier as its average value across all accounts in that category. Table 4 presents the per category accuracy for Naive Bayes and Support Vector Machine for all three cases. This table reveals that In all cases, certain categories

show higher accuracy. There are categories with higher number of LoA2+/aligned tweets such as basketball and soccer. Furthermore, accuracy for Case 1 is higher than Case 2 and Case 2 is higher than Case 3 which means better training, results in more reliable classification. Another general trend in this table is that SVM outperforms NB in Case 1 and Case 2 but in Case 3 NB surpasses SVM which can be explained by the size of dataset. Since Naive Bayes is a generative classifier it is trained better with larger dataset.

The most interesting point that we learn is that there is a relationship between accuracy and LoA2+/aligned metric that we defined in Section II. This relationship is depicted in Figure 8. This figure is a scatter plot of aggregate accuracy versus LoA2+/aligned for all categories. As this figure reveals higher number of LoA2+/aligned is equivalent to higher accuracy and vice versa which is consistent with our hypothesis. We selected LoA2+/aligned because it is the most informative feature according to our decision tree.

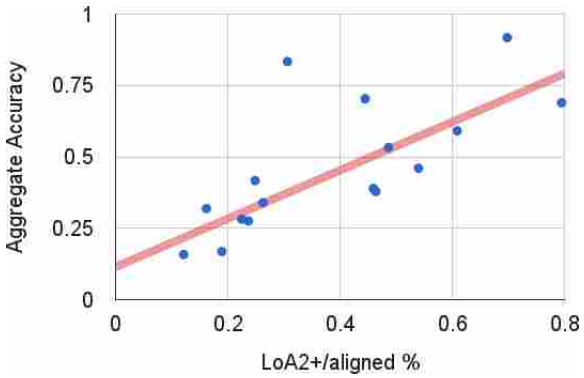
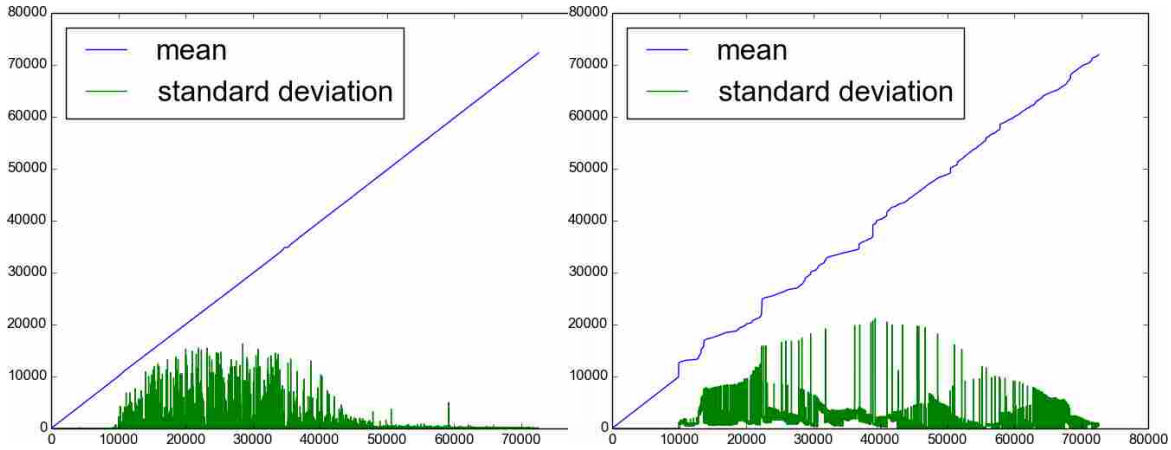


Figure 8. Scatter plot of aggregate accuracy versus LoA2+/aligned for all categories

Per Account Analysis. In this section, we focus on the accuracy of classifiers in each scenario for individual accounts. Toward this end, we plot the accuracy of SVM classifier in a heat map where X axis presents the accounts list



(a) Support Vector Machine (b) Naive Bayes

Figure 9. Average and standard deviation for all 70k values across all rounds

(accounts are grouped based on their category) and Y axis shows the category. Each cell (i, j) shows how often account j's tweets are classified as i. The bluer the cell the less accuracy and vise versa. Figure 7 shows account based accuracy heat map for SVM running on Case 1 dataset. Generally we expect each account is classified as its expected category and the diagonal red band reveals this fact, although there exist misclassification that we explain shortly.

Using the heat map, we can also visualize overlap that we discussed in Section II. Overlap between news/politics and news/finance is clearly visible that confirms our decision tree classification result that is based on LoAi/x features. We also understand from lighter vertical band above news category (13th column) that news has overlap with almost all categories.

Another interesting point here is that telecom and beverage are not classified precisely, and if we zoom in we observe that some of the low accuracy accounts are those that were aimless which approves our hypothesis in labeling section. A good example here is account VerizonWireless, which is expected to be a telecom

account while it is classified as both telecom and electronics. This is consistent with our previous findings in feature classification where electronics and telecom were classified in the same leaves although very inaccurately and also in overlap graph in presented in Figure 3.

Figure 10 plots the scatter plot between accuracy and LoA2+/aligned for all accounts which is even more revealing than Figure 8 in visualizing the relationship between accuracy and LoA2+/aligned.

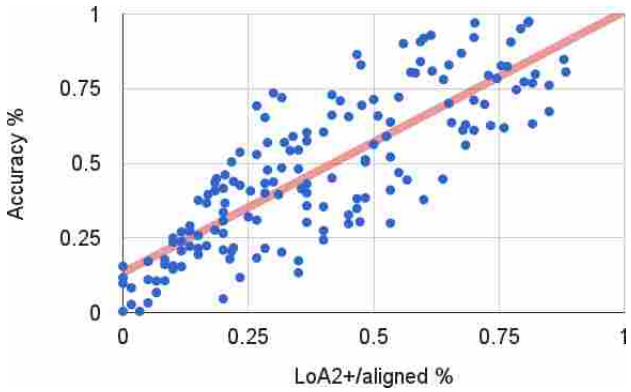


Figure 10. Scatter plot of aggregate accuracy versus LoA2+/aligned for all categories

Now that we can assign a topic to each Twitter account, we examine which keywords play the main role in inferring that topic and figure out if they are distinctive enough to separate one category from another. This analysis is done in the next section. For the next section we just consider Case 1.

Extracting Keywords

The purpose of this section is to determine the main key words that classifiers identify as distinguishing category among these collection of categories. For this analysis in addition to removing stop words we also remove URLs so

that we do not see http or https as an important keyword. After filtering we have roughly 70k keywords that may have different weights/ranks in different rounds. Therefor, first we examine the stability of keyword ranks among 70K individual keywords. In other words we are seeking to answer the following question: How consistent is the rank/weight of keywords in different rounds? For this purpose we sort all keywords in all 173 rounds and keep their ranks so each keyword has 173 ranks. Then we remove the top 35 and bottom 35 (to remove outliers). Then we compute the average and standard deviation of remaining 100 values (ranks) and plot those values for all 70k keywords.

Figure 9 illustrates this stability. It shows both average and standard deviation and apparently for the first 10k keywords the standard deviation is negligible and the average value is pretty stable, and overall SVM is much more stable than NB, which can be explained by the nature of these two classifiers because NB is a generative classifier and can not capture dependency as opposed to discriminative classifiers (e.g.SVM) that learn the boundary between classes instead of learning each class and determining as to which class each tweet belongs to. Consequently in each round Naive Bayes learns the whole data, so it produces more variable weights and consequently more variable ranks.

As a result of the above exercise we can show the keywords in a word cloud so we could visualize the words that a classifier considers important. Thus in each round of leave-one-out cross validation we sort all keywords based on their weight in a list (note that weight range is different for different classifiers since they use different algorithms to calculate weight vector hence we work with ranks instead of weight) and pick the first 200 keywords for each category. Then we plot a word

Figure 11. Top 200 keywords for category basketball – the classifier is SVM

cloud per category per classifier to visualize the keywords. A sample of these word clouds is illustrated in Figure 11 (you may find the rest of them online). In this figure, size is related to weight (but not color and centrality).

Topic Inference Through Topic Modeling

The number of topics (T) is an input for the topic modeling algorithm, and the result of this algorithm is highly dependent on this variable. In our experiment we set $T = 16$. Accordingly after running this algorithm it returns a list of 16 topics (i.e.t_0 to t_{15}) and a list of keywords associated to each topic and a mapping between documents (i.e.tweet) and topics. To present the result of our experiment we do the following exercise: Tweets of each category can be mapped to several t_i. We count how often each category is mapped to each t_i and plot the result in a heat map. Figure 12 illustrates this heat map.

As it is seen, there are certain topics that are modeled successfully, but not all of them. Despite its incompleteness, this heat map is consistent with Figure 7 in which basketball, soccer, fashion, health, and politics had relatively high accuracies.

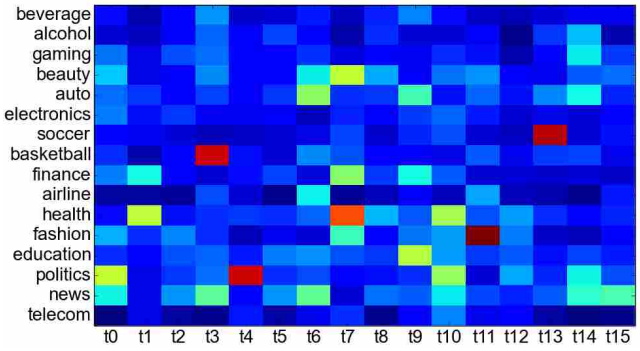


Figure 12. heat map between topic modeling result and account category

Discussion

So far we have analyzed tweets of major accounts using two methods; first we characterized tweets and extracted features (i.e.LoAi/x) and performed classification using those features. Then we feed tweets to support vector machine to obtain the accuracy. As a result of these two analysis we can think of an approach to build a valuable training set for certain applications. The approach is as follows:

- To find topic of tweets we need a labeled dataset to train the classifier.
- We measure LoAi/x features for a particular account and compare them with our result.
- If according to our division it is a purposeful account then all tweets of that account could be used for training.

Conclusion

We conducted this study in two parts, in part one we characterized tweets based on their labels and introduced a metric called LoAi/x and following is the summary of our findings:

- A majority of tweets of certain categories have an aligned topic.
- Misaligned tweets appear to be caused by multi-topic tweets that suggests pairwise relevance of topics.
- Fraction of tweets with various level of alignment offer valuable features to identify a category.
- These features also seem to reveal the way that entities in each category use Twitter.

In second part we performed text based classification and we found interesting connection between results of part one and part two:

- Certain categories/accounts exhibit higher accuracy in all cases. (e.g.soccer, basketball) these categories/accounts have a relatively higher fraction of aligned tweets (LoA2+/aligned).
- Accuracy of classification depends on the quality and the size of training dataset. More reliable training set results in higher accuracy.
- SVM outperforms NB except when we have larger data set with lower quality/reliability.

CHAPTER III

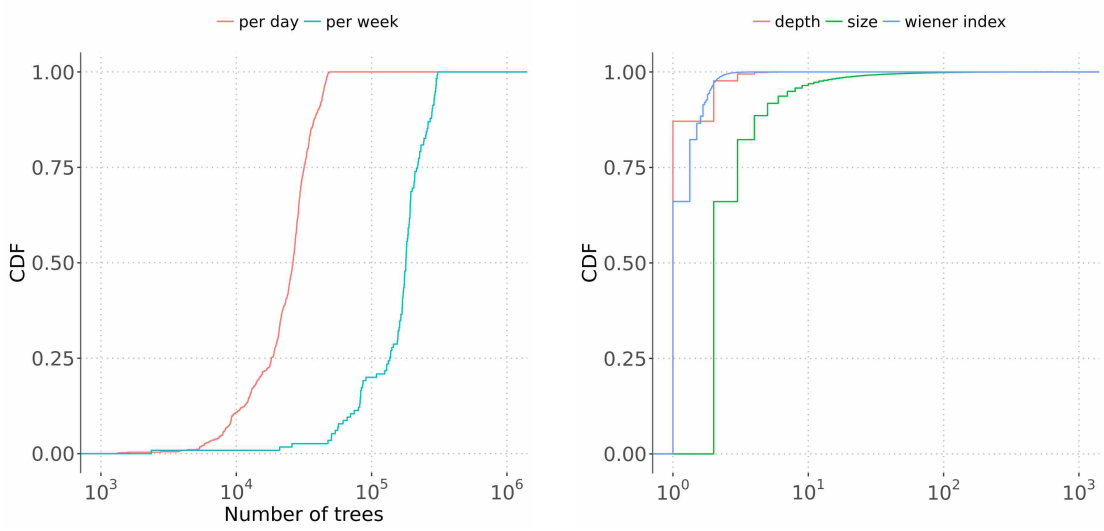
CONTENT PROPAGATION IN ONLINE SOCIAL NETWORKS

Introduction

Message propagation is a result of decision by individual users to push or pull a particular message through their social and non-social links, respectively. Most prior studies have focused on social links or simply assumed all the links in a tree are social but in practice non-social links can also be used to diffuse information. Furthermore, the importance and nature of such propagation primarily depends on whether it is relayed by (and thus informed) different groups of unrelated users or a collection of tightly related. There are SPAM trees that are artificially formed by spammers rather than the uncoordinated behaviour of users that are generally difficult to distinguish and could introduce error/noise to any such analysis.

In contrast, those messages that are relayed/reposted by a number of users (retweets, reposts) are of special interest as they engage many users beyond the followers of the initial producer of the message. This has motivated computer, social and data scientist to capture and characterize the spatial and temporal characteristics of the propagation tree for these popular messages.

A majority of these studies focus on characterizing and modeling the propagation behavior of individual trees using captured data from actual OSNs. A commonly reported finding is the skewed distribution of size and depth of these trees. This in turn implies that a majority of these trees are small and shallow. Furthermore, various studies have also pointed out that some trees are associated



(a) Distribution of number of trees daily and weekly

(b) Distribution of height, Wiener index, and size

Figure 13. Basic characterization of all trees

with spammers rather than uncoordinated relaying of a post by a group of (likely unrelated) users. it is generally not trivial to reliably separate these spam-related trees from others as they may exhibit similar characteristics. The large fraction of small (and less important) trees along with those generated by spammers could significantly affect any characterization or modeling of individual trees.

These propagation trees are often associated with a message/topic/purpose/event that is of interest to a number of users. Intuitively, the topic/purpose of a number of such trees would be related. It is valuable to determine the association among different propagation trees in order to infer more subtle patterns in information propagation beyond individual trees. Unfortunately, characterizing individual trees (along with the presence of spam users) makes it impossible to translate any characterization of individual trees to patterns across multiple trees.

A community is a collection of tightly connected (related) nodes. Therefore, the relative position of a propagation tree over the community-level view of the social graph, and the role of non-social links along with their relative position in the tree.

The goal of this study is to address two key questions: 1) how does the characteristics of propagation trees varies with respect to diffusion of content across multiple communities?, and 2) what role does the non-social links play for each group of trees (from question 1)?

Related Work

Word-of-mouth (WOM) communication is a well studied phenomenon in the literature, and content propagation in Online Social Networks (OSNs) is one of the forms of WOM mechanism that have been prevalent in recent years specially with the widespread surge of online communities and online social networks Brown and Reingen (1987) and Rodrigues, Benevenuto, Cha, Gummadi and Almeida (2011)

Here we discuss related work in several categories since information propagation in OSNs is a broad field of research, and different tracks of study are involved in it.

Characterization: Characterizing information diffusion is a very common track where a major OSN is investigated and characterized to find correlations and patterns to explain propagation. Plenty of works have been done in characterization of various well-known OSNs. Flicker is one one of the first OSNs that drew attention Cha, Mislove, Adams and Gummadi (2008), Cha, Mislove and Gummadi (2009) and Yu and Fei (2009). Twitter, thanks to its

public nature and straightforward API, is also popular among researcher Kwak, Lee, Park and Moon (2010), Cha, Benevenuto, Haddadi and Gummadi (2012) are focusing on propagation of news in Twitter, Lerman and Ghosh (2010) is another work that studies spread of news on Twitter and Digg, Ottoni et al. (2014) is a cross OSN study investigating how users retweet and repin on Twitter and Pinterest, respectively. Facebook researchers in Dow, Adamic and Friggeri (2013) and Sun, Rosenn, Marlow and Lento (2009) study the large cascades on Facebook. Reddit, which is a platform supporting online communities, is the subject of characterization in Choi et al. (2015). They study conversation patterns in terms of volume, responsiveness, and virality. Generally, these papers lack the insight necessary for investigating such a phenomenon since they treat all nodes/edges the same while different connections, users have different roles/importance. For example degree of a node alone may not reveal its importance.

Modeling and Predicting: Modeling and predicting cascades is a popular line of work in the area of information propagation. The common term is usually used in this area is information diffusion. In this approach content spreading is described using the activation process. A node could be either activated meaning it has received the information or inactive and ready to get activated with a certain probability. Thus, the propagation process is defined as consecutive activation of nodes in the network Kempe, Kleinberg and Tardos (2003). This model that is based on independent individuals who affect their neighbors is called *Information Cascade* Goldenberg, Libai and Muller (2001). Another widely used model is Linear Threshold in which each user u is influenced by its neighbor v by a certain threshold $t_{u,v}$ Granovetter (1978). These models are usually applied on the social network where there is no sharing information available. However, there exist

modeling studies that try to define a prediction problem and solve that using machine learning approaches. Authors in Cheng, Adamic, Dow, Kleinberg and Leskovec (2014) ask the question of "C*an cascades be predicted?*" and after showing that it is difficult problem Weng, Menczer and Ahn (2013), they define a problem of cascade growth in which the problem is reduced to: "*given a cascade that currently has size k, predict whether it grow beyond the median size $f(k)$*" Cheng et al. (2014). The problem definition in this approach is interesting and the result is promising however it has a drawbackin order to predict whether a cascade with size k will reach its median size they have to observe at least first k reshares which makes the problem less attractive since the goal is to find characteristics of a viral content (photo in this study) while in this work they content should already propagated k times.

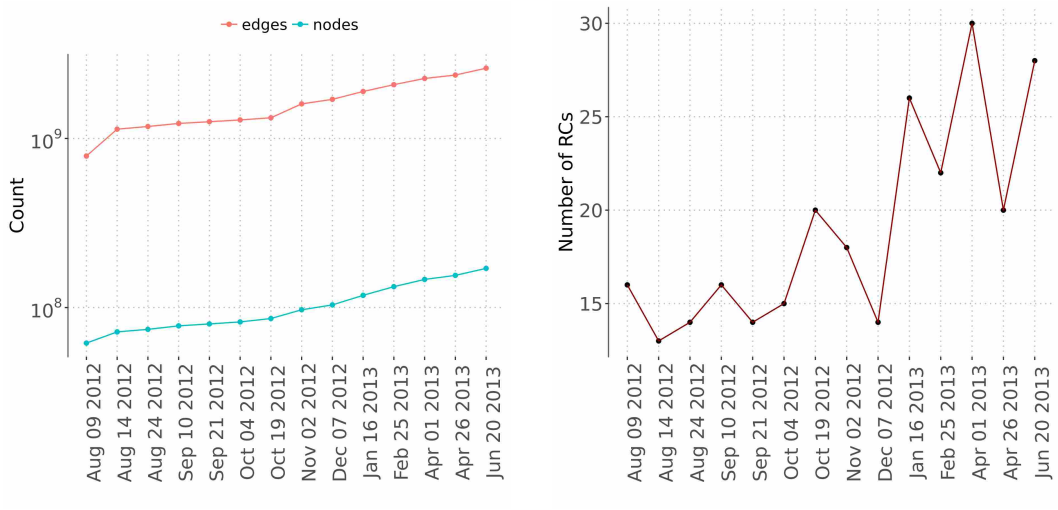
Influence: Finding and targeting influentials in Online Social Networks is another important field of study that benefits many applications such as politics, sport, and above all, marketing. Brown and Reingen (1987) is one of the earliest works in this area that claims word-of-mouth communication (WOM) is the most important source of influence. However with the advent of online social networks WOM has been replaced by social links where parameters such as degree (number of followers and/or followings), retweets, replies, mentions, and presence of URLs are leveraged to quantify social influence Cha, Haddadi, Benevenuto and Gummadi (2010), Ye and Wu (2010), and Bakshy, Hofman, Mason and Watts (2011). Finding influentials is sometimes dealing with clustering where based on a definition an optimization problem is formed and is maximized to satisfy the definition. This problem is called *influence maximization* and first addressed in Domingos and Richardson (2001). For example, Saito, Kimura, Ohara and Motoda (2016) defines

influentials as "*nodes which, if removed, decrease information spread*". Basically, they maximize the difference in the amount of *influence degree* as a result of individual node removal. Borrowing from influence cascade model, authors in Kempe et al. (2003) define their *influence maximization* problem as "finding k vertices in the graph such that under the influence cascade model, the expected number of vertices influenced by the k seeds is the largest possible". However, in this work, we are seeking to find influentials in a data-driven manner since we have all sharing information.

Google+: Content sharing in Google+ is only investigated in Kairam, Brzozowski, Huffaker and Chi (2012) by researchers from Google. They explore *Selective Sharing* in Google+ and study how active users select their audience. In fact, they investigate private sharing in Google+. There is another work by Google about Ripple visualization that is not in the are of information dissemination. In terms of OSN characterization, Gonzalez, Cuevas, Motamedi, Rejaie and Cuevas (2013) is a study during the first year of Google+ operation but it does not cover content propagation in Google+. Hence our work and our dataset is unique in terms of OSN, scale and approach.

Dataset

We have collected all public posts of all users in the Largest Connected Component of Google+ from June 28, 2011 to July 3, 2013. The number of activities (posts) in our dataset is roughly 540M. Along with the content of the activity we also retrieved attachment type, number of reshares (public + private), number of plusones, and number of replies. We refer to this dataset as *activity*.



(a) Snapshot size in terms of node and edge count

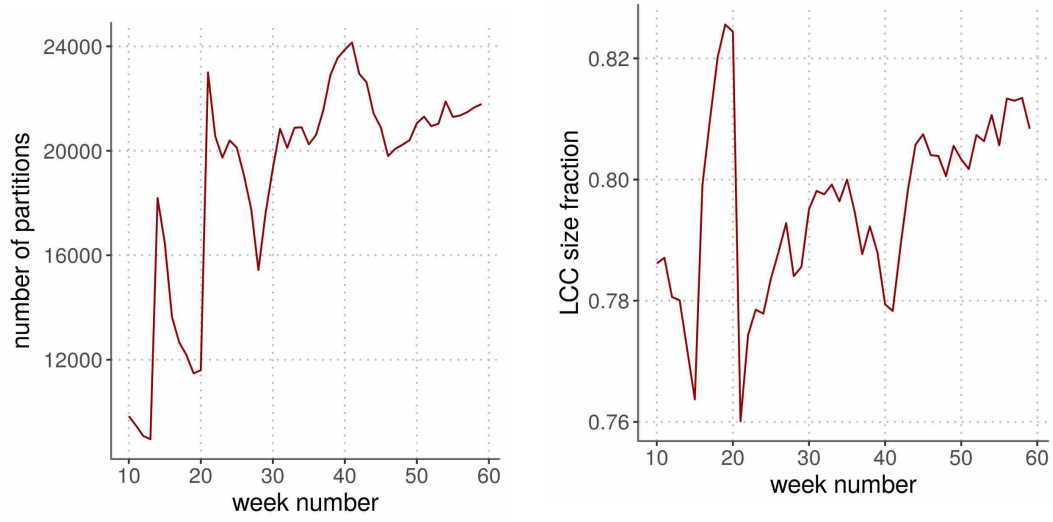
(b) Number of Resilient Communities per snapshot

Figure 14. Snapshot characteristics

Having above information, we cannot create a propagation tree in which the message originator is the root, users who reshare the original message are vertices and each sharing activity is an edge. So, we need more explicit resharing information to build the trees. This is where *Ripple* comes into play.

Ripple is a data visualization graph built-in to Google+, and is enabled when a user reshares a post publicly (one edge per reshare is added to the tree)[1]. For Ripples we obtained 29.6M reshare trees that include 90M nodes (this is the number of activities associated with ripples) and 6.5M unique users. This dataset was collected from June 17, 2011 to September 9, 2013. We refer to this dataset as *ripples*. Figure 13a shows the distribution of number of trees daily in weekly. As this figure reveals there are roughly 50K and 300K ripples daily and weekly, respectively. Also Figure 13b illustrates the distribution of size, height, and wiener

[1]As of May 20, 2015 the Ripples feature in Google+ is no longer available



(a) Each Ripple Graph consists of a large number of small partitions

(b) Largest Connected Component constitutes at least %70 of the whole graph

Figure 15. Ripple Graph Partitioning and LCC information

index[2] for all trees. This figure reveals that more than %55 of trees have size 2. These trees represent activities that are reshared only once. It also shows that more than %83 of them do not go further than one level. These contents may get propagated several times and be more viral than those which receive only one reshare but their corresponding trees are not deep. Thus we should find a technique to filter out those that do not contribute in overall content propagation across Google+. Filtering naively based on threshold on the size or height of trees is not a good idea as it only removes small and shallow trees while does not affect SPAM trees.

Note that from *activities* dataset we know how many times a post has been reshared in total (publicly and privately) but from tree size obtained from *ripples* dataset we can only find number of public reshares! Apparently we do not have

[2]Wiener index is a measure for virality such that trees with a low Wiener index resemble star graphs, while those with a high index appear more viral.

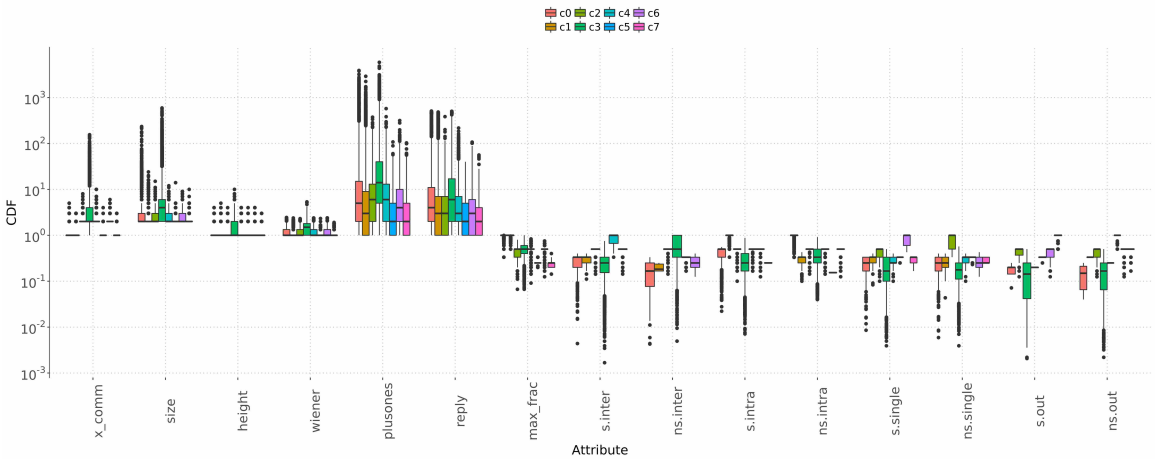


Figure 16. distribution of all attributes for eight clusters

private propagation information. Furthermore, if a user reshares her own content several times it does not show up in the number of reshares (i.e.duplicate reshares are eliminated from number of reshares field) hence these two numbers are not the same essentially. Although we can understand what percentage of reshares is private.

The third dataset addresses connectivity information of users across Google+. We have access to 14 snapshots of Google+ structure that are crawled roughly one month apart, starting August 2012 to May 2013. Each snapshot has a directed edge view in the form of $E = (v, w), v$ follows w. The network size ranges from 60 million nodes in the first snapshot to around 160 million nodes in the last one. The number of edges varies in the range of 800M edges in the first snapshot up to 2.6 billion edges in the last one. In all snapshots, average degree fluctuates between 30 and 40. We refer to this dataset as *connectivity* dataset. Figure 14a illustrates the number of nodes and edges across all 14 snapshots and we observe an upward trend in snapshot size.

Since the goal of this study is to investigate the relative position of a propagation tree over the community-level view of the social graph we use community detection technique to obtain the communities across social graph. A community is a collection of tightly connected (related) nodes. We focus on the subgraphs of high degree nodes, i.e.*core nodes*, and identify the community of core nodes, i.e.*core communities*, as the main elements of the graph. Most community detection methods are non-deterministic that results in community mapping variation. To minimize this effect, we run the community detection technique on the core subgraph n times. Then, we compare the communities that each node were mapped to and identify groups of nodes that have identical mapping vectors. We refer to each group of core nodes as *resilient communities*. The main tuning parameter for this approach is the number of high degree nodes, i.e.size of the core subgraph. This could possibly change the number of communities, and therefore change the resolution of our view. We refer to each core subgraph as a *view* and consider top five thousand most followed nodes. We plot the number of Resilient Communities (RC) per snapshot in Figure 14b and we see a correlation between node/edge count and number of RCs. In terms of size, RCs usually span from ten nodes to 1,000 nodes and they rarely get bigger.

If we cross trees and communities we notice that all communities have at least one tree crossing them. Furthermore, %99, %96, and %70 of communities have at least two, five, and ten trees crossing them, respectively. This implies that conversations do happen among communities and the question is do they cross communities?, remain inside communities?, or connect different communities? Another parameter to consider is whether tree edges are social (tree edge is present in the social graph or a user reshares from one of its neighbors) or non-social (it

happens when a user reshares from a user who is not directly connected to) and in each case where they are located. Are they located between nodes in different communities or between nodes in the same community. To answer to the former question we need to dig deeper but for a quick response to the latter we define edge type as follows (note that we are focusing on tree edges and we refer to Resilient Community as community for brevity):

- intra-community edges: the two sides of an edge are in one community.
- inter-community edges: each side of an edge belongs two a separate community.
- single-community: one side of an edge belongs to a community and another side does not belong to any community.
- out-of-community: none of the sides are present in any community.

If we also consider whether an edge is social or non-social, we would end up with eight edge types: social intra-community, non-social intra community, etc and here are some basic stats about the location of social and non-social links:

	social	non-social
intra-community	%36	%64
inter-community	%35	%65
single-community	%39	%61
out-of-community	%39	%61

Table 5. Location of social and non-social links with regards to communities

Table 5 presents the relative location of social and non social edges with regard to communities. One important phenomenon, which is somehow counter-

intuitive, is that a great fraction of links are non-social! This implies that social connection is not a significant factor in information propagation in Google+. In fact, Google+ is a social layer across all services provided by Google and the users' timeline is fed from different sources including YouTube, Google search engine, Google+ recommendations, etc. Hence, it is not surprising that users share content from users who do not follow. Another important observation is that edges that are less affiliated with communities are more social. This suggests that community is a factor in information diffusion. According to the location of non-social links, we can argue that it is more likely for the users in a community to share each other's content despite being socially disconnected. Another question that will arise here is: what is the social distance between two sides of non-social links? If this distance is small then we can argue that we can predict that it is very likely that such a link can form and becomes social. In order to investigate this we can find non-social links that take part in propagation in one week and study the link formation at a later time. (This could be defined as a prediction problem)

cluster number	largest fraction of edge type	sociality	size	height	popularity	maximum overlap with a community	number of crossed communities	number of trees in the cluster
cluster0	intra-community	social	moderate	shallow	high	very large	very small	22.9K
cluster1	intra-community	non-social	small	shallow	low	very large	very small	21.1K
cluster4	inter-community	social	moderate	shallow	moderate	moderate	large	8.0K
cluster3	inter-community	non-social	largest	largest	highest	moderate	largest	17.3K
cluster6	single-community	social	moderate	shallow	moderate	large	large	4.0K
cluster2	single-community	non-social	moderate	shallow	moderate	small	large	4.5K
cluster7	out-of-community	social	small	shallow	unpopular	smallest	smallest	1.7K
cluster5	out-of-community	non-social	small	shallow	unpopular	small	small	1.5K

Table 6. Summary of eight clusters

Methodology

Rather than individual trees, we consider an aggregate view of a group of trees that occur within a window of time that form a directed graph called Ripple

Cluster type	community set	**CS0**							
social intra-community	**CS0**	4,361	**CS1**						
non-social intra-community	**CS1**	2,494	4,113	**CS2**					
non-social single_community	**CS2**	1,253	1,070	1,726	**CS3**				
non-social inter-community	**CS3**	3,906	3,614	1,588	6,205	**CS4**			
social inter-community	**CS4**	2,238	2,403	1,128	3,206	3,553	**CS5**		
non-social out-of-community	**CS5**	27	22	22	33	22	33	**CS6**	
social single-community	**CS6**	1,278	1,380	759	1,820	1,402	25	1,988	**CS7**
social out-of-community	**CS7**	26	28	20	31	26	5	23	34

Table 7. number of communities in each Community Set and overlap between pairs of CSs

Graphe or RG. Ripples span over 116 weeks so for each week we can generate a graph which is a union of all ripple trees occurring during that week. We group all ripple trees in a weekly basis using tree time-stamps and superimpose them on each other to create a weekly graph. The resulting graph is a weighted, directed graph and we can also determine if an edge has a social tie or not. Therefor we would end up with a rich graph from which we could extract useful user properties including measures for user centrality in the discussion network in addition to the social network. aggregation within a window of time is motivated by the fact that related events are more likely to occur in a closely related window of time. We examine different length for the aggregation window and show that the characteristics of RG is not sensitive to the duration of window, we choose one week and call the graph WRG.

Characterizing WRG shows that it has a large connected components that contains a large fraction of all trees, other trees are considered less relevant. Since we are looking for trees that are big and possibly have overlap with many other trees we eliminate trees that are not in the Largest Connected Component

(LCC) of a WRG. Note that we are removing trees that do not take part in the main stream of conversations this is roughly %30 of the WRG. Figure 15a which illustrates the number of partitions per WRG reveals that there are many partitions in a WRG but Figure 15b shows that the LCC covers more than %70 of the WRG which is a remarkable fraction.

Next, we identify all communities in connectivity snapshots, then detect Resilient Communities to avoid uncertainty in community detection. Then we layout individual trees in the LCC of WRG over the community level view of the connectivity snapshot. We run clustering across all trees based on the relationship of trees and communities as well as characteristics of individual trees (e.g.size, height, Wiener Index) to see whether there are some clear distinction between different trees. this reveals different types of trees with respect to propagation across (or within) communities. Finally, we characterize communities based on the characteristics of their crossing trees.

Tree clustering

The goal of this section is to see whether trees exhibit different characteristics with respect to the mapping through different communities and also examine the role of non-social links. To this end, we first perform clustering on trees of LCC of WRG considering different sets of features and then explain each cluster in details. The main set of attributes that we are interested in is the fraction of different edge-types in trees with respect to communities. Thus we consider following 10 attributes for clustering algorithm:

(a) cluster 0 (b) cluster 1 (c) cluster 2

(d) cluster 5 (e) cluster 4 (f) cluster 5

(g) cluster 6 (h) cluster 7

Figure 17. content analysis

Attributes that explain the relationship between trees and communities (6 attributes):

- fraction of edges that are intra-community
- fraction of edges that are inter-community
- fraction of edges that are single-community
- fraction of edges that are out-of-community
- number of crossed communities
- maximum fraction of the tree in a community

Attributes related to each individual tree (4 attributes):

- fraction of edges that are social/non-social
- tree size
- tree height
- tree wiener index

We run k-means algorithm (all features normalized) on all trees with respect to the listed features and we find that trees are nicely clustered into eight groups.

Figure 16 illustrates the summary distribution of attributes in each cluster. Note that the distributions of some other features that did not take part in clustering are also plotted to have a clearer view about cluster distinction. Table 6 presents the summary of these 8 clusters. We can understand from this table that trees in each cluster have a large fraction of one specific edge type. For example trees in cluster0 consists of edges that are social and inside one community. Table 6 also reveals that very large and popular trees (cluster3) usually cross many communities and many non-social links are involved in the propagation of these types of content. On the other hand, the most popular social trees are contained in a single community and even though they are large they do not go very deep, i.e.their shape is like a star.

In terms of content of the ripples that are present in different clusters we cannot see any clear distinction. Figure

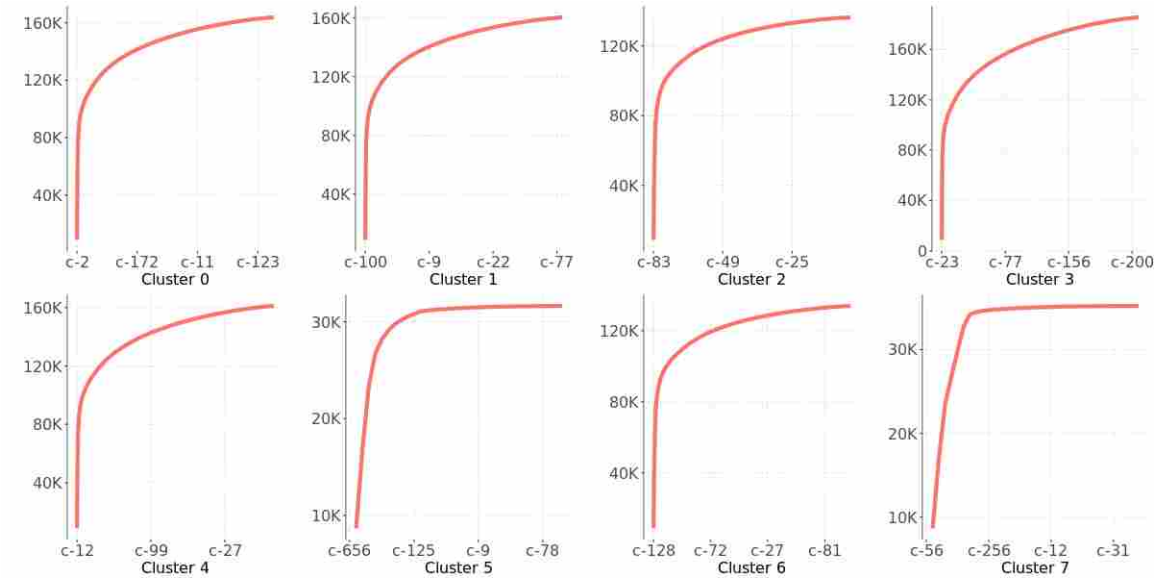
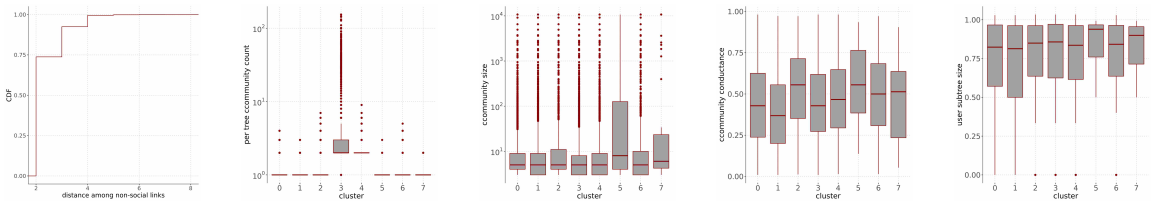


Figure 18. tree locality per cluster

Next we show the locality of trees across communities. We consider all trees in a cluster, obtain the union of all communities that these trees cross, for each community count the number of unique trees that cross the community, then sort the communities based on the number of crossing trees, and then plot the number of unique trees that cross communities. Figure 18 shows tree locality per cluster.

To further examine the role of non-social links we calculate the shortest distance (through the social graph) between connected nodes by non-social links in each group of trees. Figure 19a illustrates the distribution of pairwise distance among social links. As this figure shows %75 of edges have distance of 2 which means they are one node away from being connected. This suggests that users tend to share content not only from their immediate connections but also from their two hop neighbor-ship and such links have the potential to form.

Extreme cases. Large trees contained in one community: if we filter trees based on number of community they cross (#crossed community



(a) (b) (c) (d) (e)

Figure 19. (a) pairwise distance among non-social links, (b) community count per tree per cluster, (c) community size per Community Set, (d) conductance per Community Set, (e) subtree size per Community Set

= 1) and tree size (tree size $>$ 20) we will end up with 38 trees that are all in cluster 3. As we inspected these trees manually, they mainly belong to non-English communities. The majority of them are Chinese and the rest are Japanese or Arabic. To explain this phenomena we can argue that the users that are present in these communities are tightly connected and their contents do not go beyond their community. In other words these communities are bounded by language.

Small trees that span across many communities: in this case we consider trees that cross as many community as their size and we select those with size less than 6 and greater than 2. Focusing on these trees we notice that more than %70 of them have depth 1 meaning that they are star-like trees. It implies that there are contents that are disseminated through multiple communities by one hop only and then they fade away perhaps because of other aspects of social structure. To study this behavior we examined ***border edges*** (i.e.edges whose sides fall into two different communities), and we realized that half of them are social. This is larger than %30 social fraction among trees in the LCC of WRG, and we can argue that a content propagates outside a community because of a social edge that is not captured by Luvain algorithm due to weaker attachment to the community (smaller modularity).

Community Level Analysis (per cluster)

In this section we investigate the roles of communities in propagation. First we calculate number of communities that each tree crosses per cluster. Figure 19b illustrates the summary distribution of number of communities each tree crosses per cluster. We observe that clusters 3 and 4 that include trees with larger fraction of inter-community edges contain more unique communities. This makes sense since trees that connect more communities have more inter-community edges. In other clusters trees mainly cross one community.

We identify the set of communities that all trees in each cluster cross. Thus, we will have eight set of communities (CS0 to CS7), and these sets may have overlapping communities. The number of communities that exists in each set and the number of overlapping communities among this Community Sets is presented in Table 7. The highlighted diagonal shows the number of communities in each CS. We see that CS5 and CS7 include very small number of communities. These community sets belong to clusters with trees that have a large fraction of *out-of-community* edges. CS3 (non-social inter-community) contains the largest number of communities with 6,205 communities and generally it has a large overlap with other CSs. Figures 19c and 19d depicts different characteristics of these community sets. From Figure 19c we understand that community sets 5 and 7 that are crossed by cluster of trees with large fraction of out-of-community edges are larger. This means communities that are crossed by trees which have many out-of-community edges are large. This makes sense because these kinds of trees are usually isolated trees and they may touch large communities as those communities are more scattered across network and the likelihood of crossing gets bigger.

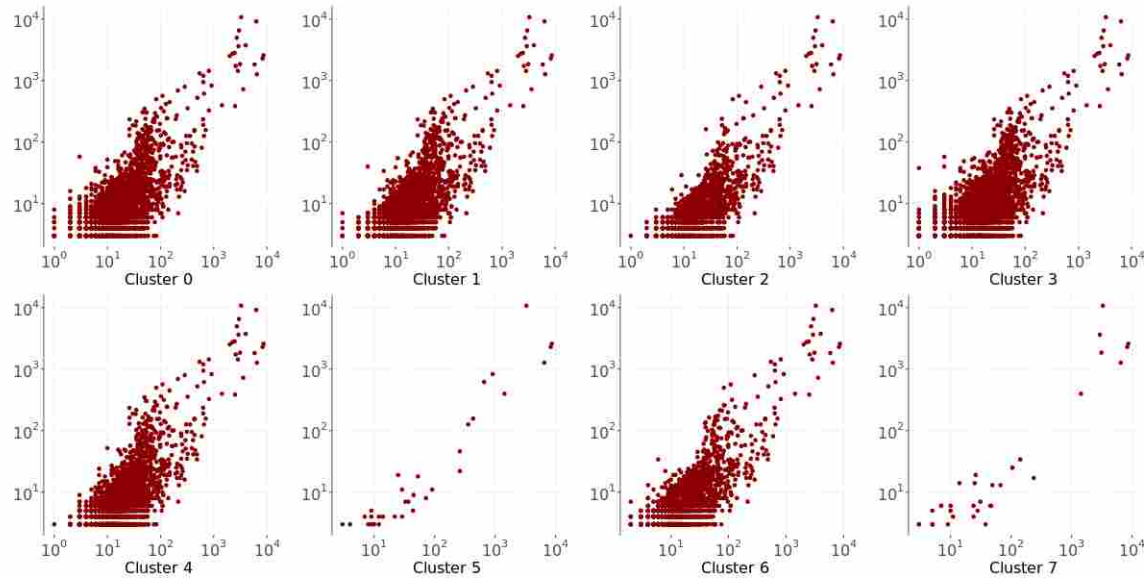


Figure 20. Community size versus tree count per cluster

Furthermore, we plot community conductance per cluster in Figure 19d. A higher conductance score means that a set of nodes more closely resembles the connectivity pattern of a community. It shows that communities in cluster type of single-community and out-of-community have more conductance score. We can bring the same argument as to why communities in CS5 and CS7 are larger. The isolated trees that cross these communities have limited connection with the rest of network and as this figure reveals only larger and denser communities can reach them.

Next we examine the context of communities that exhibit extreme characteristics. For this purpose we need to define communities with extreme characteristics. Here are some examples of extreme communities:

1. communities that are present in all Community Sets

2. communities that only appear in CS5 and CS7 and are small(those that include trees with large fraction of out-of-community edges and are not very large)

3. communities that are small but many trees cross them.

4. communities that are large but crossed by a small number of trees.

5. communities that generate content

6. communities that consume content

To obtain the context of a community we can simply obtain the topic of crossing trees by topic modeling. However, this is not a good approach since the overlap between a tree and a community could be just one user and the context of tree is not related to that community. Hence we consider the content of trees whose root users are present in that community. With this approach we also take into consideration the role of trees.

case1: communities that are present in all CSs: There are only 5 communities that are present in all Cluster Sets. These communities are very large in terms of crossing trees and crossing users. There is no point in checking the context of these communities since these are just 5 very popular communities.

case2: communities that only appear in CS5 and CS7: 10 communities only appear in these two sets. The topic modeling results shows that each topic is assigned to one community. For example topic 1 which is about computer is assigned to community number 94108. This community consists of three users all connected and they all post about computer and technology.

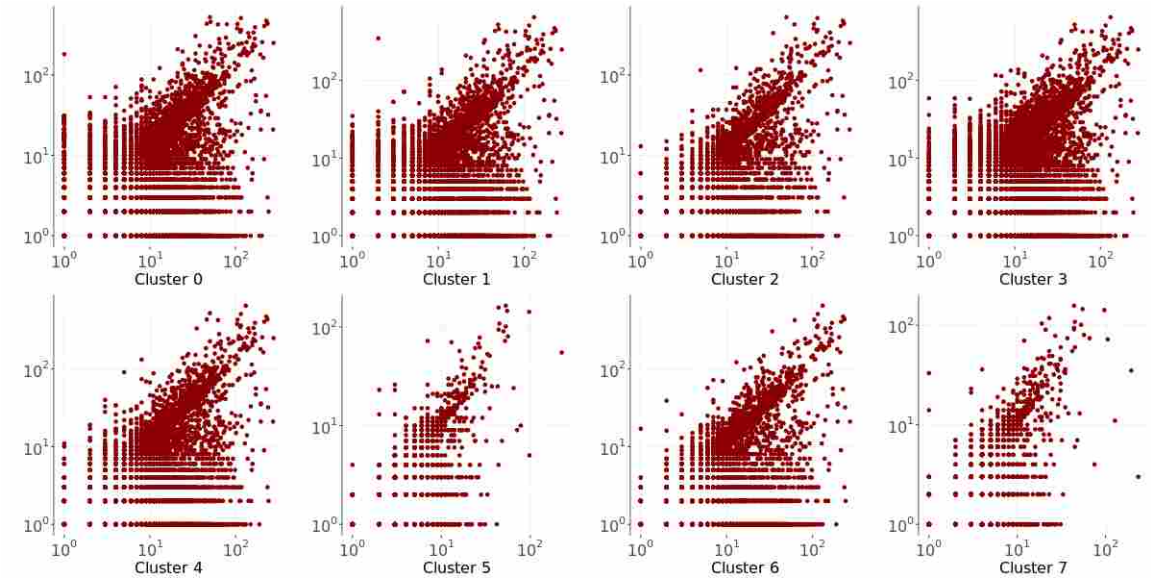
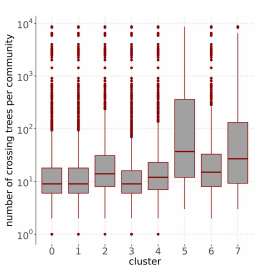


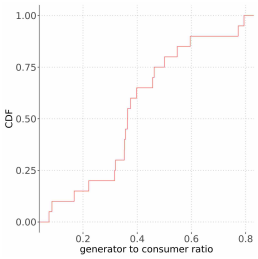
Figure 21. Node degree in WRG versus tree count per cluster

However not all topics are meaningful since these communities are small and cannot be examined contextually, necessarily.

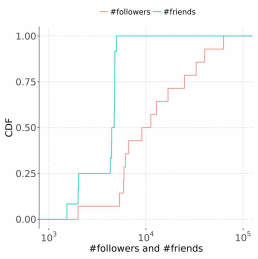
case3: communities that are small but many trees cross them: In this category we consider communities that are smaller than 10 nodes and they are crossed by more than 100 tree. considering these two criteria we find seven communities. We should note that crossing many trees does not mean all trees



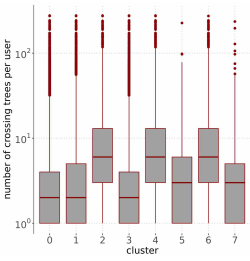
(a) number of crossing trees per community



(b) distribution of generator to consumer ratio



(c) distribution of number of followers and number of friends



(d) number of crossing trees per user

Figure 22. Locality analysis

are generated in this community. In fact very few contents are generated in these communities. As an example, community 146340 consists of 6 users who are either brands or female users interested in fashion and topic 4 (top 10 words: *exclusive, inside, click, check, jewelry, enterprise, teaser, kim, colorful, stage*) is assigned to this community correctly.

case 4: communities that are large but crossed by a small number of trees: We set number of users greater than 100 and number of crossing trees less than 40 and we would end up with 11 communities. Topic modeling shows meaningful results for these communities. For instance, content generators in community 141913 mainly post about economics. Very popular users such as "The Economist" generate content in this community and its topic is topic 2 (Top 10 words: *world, economy, new, right, technology, release, country, government, analytics, growth*) which can be labeled as economics intuitively.

case 5: communities that generate content: All users in these communities generate content. These communities are very interesting. They are very homogeneous in terms of user context. For example:

- all 13 users in Community 141871 are related to restaurants and vacations affairs.
- community 123984 consists of 8 European male bloggers.
- community 124041 posts about romance and all its users are female writers.
- all 7 users in community 132176 are female users blogging about social media. They are managers, consultants, and business women.

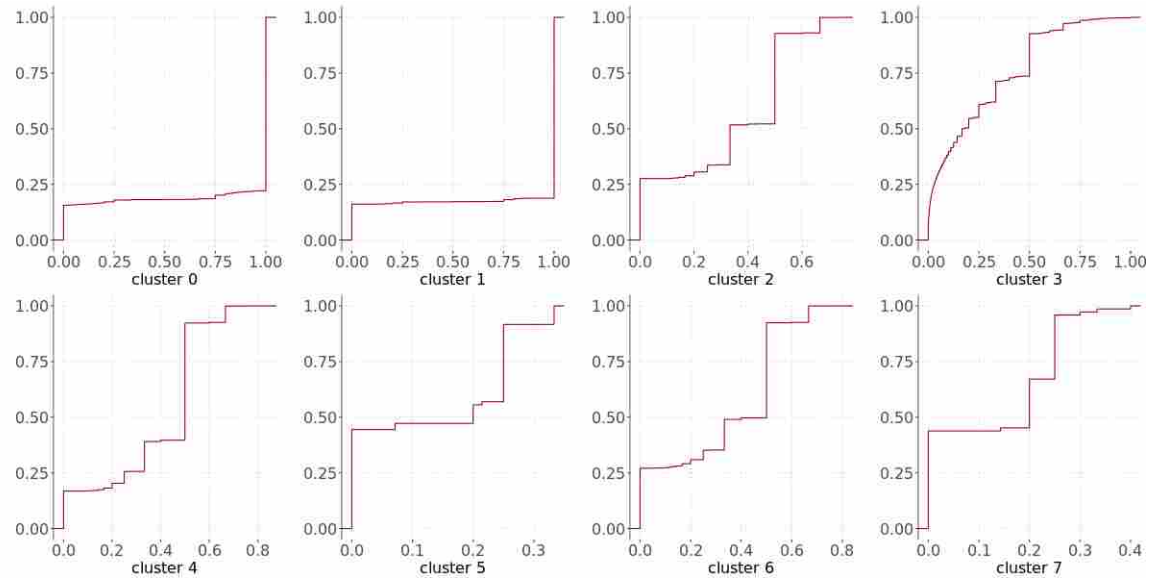


Figure 23. fraction of trees that cross each community per cluster

- users in community 127765 are Greek computer enthusiasts.
- Spanish users post in community 127513.

Note that we remove non-English content but it is apparent that non-English users generate English content as well and that is why the topic modeling algorithm finds the topic of non-English communities correctly.

case 6: communities that consume content: None of the users in these communities generate content. In this case, since there is no content generator in these communities we feed the topic modeling algorithm with the content of all users in the community. Again the main observation is that users in a community share a common theme. For example, community 130112 is an Indian/Pakistani community so is community 129729, but the content in these communities is not as coherent.

The takeaway of this exercise is two-fold. First, most communities have something in common in terms of linguistic features regardless of size or other tree-level/community-level attributes. Second, communities show distinct characteristics when it comes to content generation and content consumption.

We also calculate the average size of sub-trees that are hanging from each crossing node (in each crossing tree in a cluster). Figure 19e reveals that due to the large number of users who are present at the leaves of the trees, the values are generally small. This result is aligned with the community size per CS 19b which means the larger the community the higher the average sub tree size.

Locality Analysis

top 10 communities that have most crossing trees: Considering individual communities per cluster, for each community we capture all trees that cross that community. Figure 22a illustrates the summary distribution of the number of crossing trees per community per cluster and we see that clusters 6 and 7 has large number of trees crossing their communities. The reason is that only large communities exist in those clusters as explained earlier. Besides, Figure 20 is a scattered plot of community size vs number of crossing trees per cluster. Note that communities in different clusters has a large overlap with each other which explains the similarity among these sub figures. See also Table 7 for pairwise overlap between community sets.

Next we explore the top 10 communities with the largest number of crossing trees per cluster. Out of 10 communities per cluster, 6 communities are common among all clusters except clusters 5 and 7 (clusters containing trees with larger

fraction of out-of-community edges). The number of unique communities is 20 which are very large and have a large generator to consumer ratio (Figure 22b).

top nodes with most crossing trees per cluster: there are 26 unique users in the set of top nodes with most crossing trees per cluster. These user are almost common across all clusters but clusters 5 and 7. In terms of number of followers these users are pretty popular (Figure 22c). 20 of these users belong to 16 distinct communities, and the other 6 do not belong to any community and interestingly these 6 users have fewer number of crossing trees. See table 8 for the detail and the pointer to the user profile of these users. Note that this table is sorted based on number of crossing trees per user (third column) and user names are click-able. Last column of this table shows which clusters each user has appeared in. As this table presents, out of community users never appear on trees in cluster 3 (inter-community non-social) and some of them are not present in cluster 2 (single-community non-social).

Figure 23 depicts the fraction of trees that cross each community per cluster. As this figure shows the distributions are similar for pair of clusters such as cluster 0 (social intra-community) and cluster 1 (non-social intra-community) or cluster 2 (non-social single-community) and cluster 6 (social single-community). When the tree size is large (i.e.cluster 3) the we have different values for the fraction, however when tree size is small (i.e.cluster 0 and cluster 1) trees are contained inside one community which explains why more than %75 of the trees have complete overlap with communities.

url	comm_id	#trees	clusters
Hallo K	128262	276	0,1,2,3,4
u02	147255	273	0,1,2,3,4,5
Creative Ideas	147738	242	0,1,2,3,5
Gong Xiuzhi	147693	239	0,1,2,3,4,5
lilian michalski	145187	235	0,1,2,3,4,5
teresa itapua	146765	234	0,1,2,3,4,5
Gustavo Keive	142489	233	0,1,2,3,4,5
domi gautier	147723	227	0,1,2,3,4,5
Gary Johnson	127369	226	0,1,2,3,4,5,6
Uri Palatnik	144567	215	0,1,2,3,4,5
Jovan Vari	147723	214	0,1,2,3,4,5
Manuela Azevedo	147723	196	0,1,2,3,4,5,7
Brian Gauspohl	146975	128	0,1,2,3,4,5,7
Matt Uebel	147649	98	0,1,2,3,4,5,6
Shinji Tanaka	119727	78	0,1,2,3,4,5,6
Dawn Page	147713	75	0,1,2,3,4,5,7
Vincent Lagrandmaison	147723	75	0,1,2,3,4,5,7
Saichon Prasert	145760	71	0,1,2,5,6
Arun Bector	NA	67	0,1,2,4,5,7
Andrew King	146436	66	0,1,2,3,4,5,6
Anthony Russo	147649	65	0,1,2,3,4,5
Jose Andres	NA	64	0,1,2,4,5,6,7
Richard Ricciardelli	NA	56	0,4,5,6
HighclimbingFate	NA	55	0,1,4,5,6,7
Carlos Ramirez	NA	53	0,1,4,5,6
Aeraj Ul Haq	NA	53	0,2,4,5,6,7

Table 8. top nodes with most crossing trees per cluster

Conclusion

There are three main contributions in our work:

1. We generated a graph called WRG by superimposing all the trees that occur in one week. Using this graph we got rid of spam and captured the main stream of conversation that happens in one week.

2. We showed that trees can be grouped into eight meaningful clusters and we characterized trees in these groups and showed that each has its own characteristics w.r.t. communities.

3. We should that social links are not the main cause of propagation and communities play a significant role in relaying content. We also showed that many (more than %75) non-social links that share content from each other (are present in trees) are one social link away from each other.

APPENDIX A

THE DECISION TREE BASED ON LOAI/X FEATURES

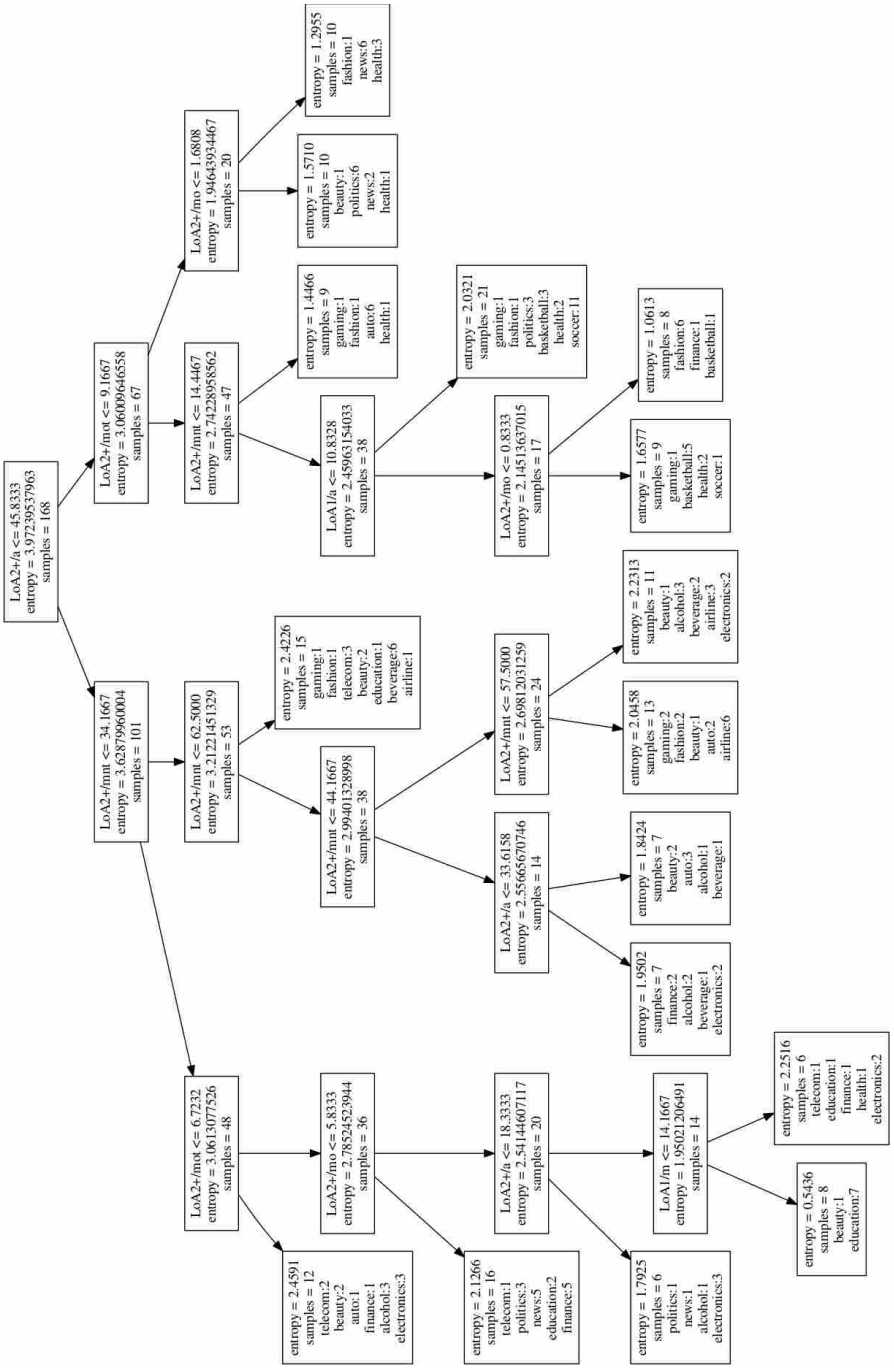


Figure A.24. The decision tree based on LoAi/x features

APPENDIX B

LIST OF ALL TOPICS WITH THEIR ASSOCIATED ACCOUNTS

Topic	Accounts associated with the topic	number of tweets per account
finance #accounts: 6 #tweets: 31,776	Bloomberg	3,233
	BofA_Community	3,198
	Citi	3,215
	NASDAQ	3,243
	Visa	3,215
	Sequoia_Capital	2,772
health #accounts: 10 #tweets: 27,726	WebMD	3,205
	MayoClinic	3,233
	EverydayHealth	3,229
	ClevelandClinic	3,238
	HopkinsClinic	3,205
	DoveMed	1,532
	pfizer	1,720
	JNJNews	3,231
	MedicalNews	3,238
	NIHClinicalCntr	1,895

Table B.1 continued from previous page

Topic	Accounts associated with the topic	number of tweets per account
soccer #accounts: 12 #tweets:38,522	Arsenal	3,200
	FIFAcom	3,238
	UEFAcom	3,202
	premierleague	3,201
	chealseafc	3,204
	FCBarcelona	3,203
	EuropaLeague	3,222
	ChampionsLeage	3,198
	LFC	3,208
	ManUtd	3,223
	MCFC	3,212
	SpursOfficial	3,211
telecommunication #accounts: 7 #tweets: 22,583	Skype	3,252
	VerizonWireless	3,205
	ATT	3,239
	cspan	3,235
	TMobile	3,207
	sprint	3,209
	VZWnews	3,236

Table B.1 continued from previous page

Topic	Accounts associated with the topic	number of tweets per account
politics #accounts: 15 #tweets: 36,923	BarackObama	3,210
	algore	1,304
	SenJohnMcCain	3,235
	billclinton	180
	newtgingerich	3,213
	MittRomney	1,400
	GOP	3,231
	FreedomWorks	3,239
	dccc	3,223
	HouseDemocrats	3,219
	LibDems	3,215
	StateDept	3,209
	OpenGov	623
	TheJusticeDept	1,215
	ObamaNews	3,207
gaming #accounts: 6 #tweets: 19,383	PlayStation	3,220
	Xbox	3,232
	NintendoAmerica	3,237
	ASTROGaming	3,237
	elgatogaming	3,222
	ScufGaming	3,235

Table B.1 continued from previous page

Topic	Accounts associated with the topic	number of tweets per account
news #accounts: 14 #tweets: 45,044	cnnbrk	3,204
	BBCBreaking	3,223
	BreakingNews	3,232
	Reuters	3,203
	AP	3,218
	ABC	3,213
	CBSNews	3,241
	nprnews	3,205
	NBCNews	3,203
	BloombergNews	3,242
	CNN	3,198
	PBS	3,212
	CNBC	3,218
	FoxNews	3,232

Table B.1 continued from previous page

Topic	Accounts associated with the topic	number of tweets per account
airline #accounts: 10 #tweets: 32,229	JetBlue	3,248
	SouthwestAir	3,231
	AmericanAir	3,208
	Delta	3,210
	VirginAmerica	3,244
	USAirways	3,202
	united	3,240
	British_Airways	3,206
	AirCanada	3,214
	VirginAtlantic	3,226
alcohol #accounts: 10 #tweets: 28,339	TopBrassVodka	3,233
	newbelgium	3,230
	dogfishbeer	3,236
	SierraNevada	3,227
	DeschutesBeer	3,237
	budlight	1,394
	MillerLite	2,156
	Budweiser	2,234
	CoorsLight	3,206
	Skinnygirl	3,186

Table B.1 continued from previous page

Topic	Accounts associated with the topic	number of tweets per account
auto #accounts: 12 #tweets: 38,589	Audi	3,220
	Lexus	3,228
	Ford	3,216
	chevrolet	3,245
	NissanUSA	3,233
	MBUSA	3,193
	Jeep	3,204
	Toyota	3,226
	JaguarUSA	3,177
	Dodge	3,199
	VW	3,207
	GM	3,204
basketball #accounts: 9 #tweets: 28,850	NBA	3,200
	usabasketball	3,176
	Lakers	3,206
	chicagobulls	3,205
	MiamiHEAT	3,227
	celtics	3,201
	Orlando_Magic	3,195
	nyknicks	3,242
	okcthunder	3,198

Table B.1 continued from previous page

Topic	Accounts associated with the topic	number of tweets per account
beauty #accounts: 10 #tweets: 32,211	COVERGIRL	3,214
	Clinique_US	3,246
	revlon	3,203
	LancomeUSA	3,197
	Dove	3,234
	LushLtd	3,236
	tartecosmetics	3,213
	DegreeWomen	3,210
	AvonInsider	3,232
	OlayUS	3,226
beverage #accounts: 10 #tweets: 32,969	pepsi	3,202
	CocaCola	3,234
	redbull	3,221
	mtn_dew	3,237
	drpepper	3,225
	Sprite	3,212
	vitaminwater	3,976
	Tropicana	3,231
	Snapple	3,203
	Lipton	3,228

Table B.1 continued from previous page

Topic	Accounts associated with the topic	number of tweets per account
	Harvard	3,201
	UOPX	3,210
	Stanford	3,203
	UniofOxford	1,611
education	Yale	3,228
#accounts: 11	Cambridge_Uni	3,221
#tweets: 33,773	TAMU	3,224
	Princeton	3,195
	OhioState	3,229
	UTAustin	3,223
	umich	3,228

Table B.1 continued from previous page

Topic	Accounts associated with the topic	number of tweets per account
electronics #accounts: 12 #tweets: 37,522	SamsunMobileUS	3,210
	BlackBerry	3,209
	intel	3,203
	Sony	3,204
	nokia	3,203
	htc	3,201
	HP	3,244
	Cisco	3,204
	nvidia	2,926
	Dell	3,206
	lenovo	3,227
	IBM	2,485

Table B.1 continued from previous page

Topic	Accounts associated with the topic	number of tweets per account
fashion #accounts: 14 #tweets: 34,837	Dior	1,005
	CHANEL	810
	delcegabbana	3,225
	VictoriaSecret	3,234
	hm	3,198
	Burberry	3,247
	YSL	178
	CalvinKlein	2,746
	armani	3,201
	Versace	3,012
	gucci	2,500
	RalphLauren	1,998
	TommyHilfiger	3,235
	VANS_66	3,248
finance #accounts: 4	kickstarter	3,240
	WorldBank	3,203
	AmericanExpress	3,216
	CNNMoney	3,219

Table B.1.

REFERENCES CITED

Ahmed, A., Low, Y., Aly, M., Josifovski, V. & Smola, A. J. (2011). Scalable distributed inference of dynamic user interests for behavioral targeting. In *Proceedings of the 17th acm sigkdd international conference on knowledge discovery and data mining* (pp. 114–122).

Bakshy, E., Hofman, J. M., Mason, W. A. & Watts, D. J. (2011). Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth acm international conference on web search and data mining* (pp. 65–74).

Barbieri, N., Bonchi, F. & Manco, G. (2014). Who to follow and why: Link prediction with explanations. In *Proceedings of the 20th acm sigkdd international conference on knowledge discovery and data mining* (pp. 1266–1275).

Blei, D. M. & McAuliffe, J. D. (2007). Supervised topic models. In *Nips* (Vol. 7, pp. 121–128).

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003, March). Latent dirichlet allocation. *J. Mach. Learn. Res.*, *3*, 993–1022.

Bonchi, F., Castillo, C., Gionis, A. & Jaimes, A. (2011). Social network analysis and mining for business applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *2*(3), 22.

Brown, J. J. & Reingen, P. H. (1987). Social ties and word-of-mouth referral behavior. *Journal of Consumer research*, *14*(3), 350–362.

Cha, M., Benevenuto, F., Haddadi, H. & Gummadi, K. (2012). The world of connections and information flow in twitter. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *42*(4), 991–998.

Cha, M., Haddadi, H., Benevenuto, F. & Gummadi, P. K. (2010). Measuring user influence in twitter: The million follower fallacy. *ICWSM*, *10*(10-17), 30.

Cha, M., Mislove, A., Adams, B. & Gummadi, K. P. (2008). Characterizing social cascades in flickr. In *Proceedings of the first workshop on online social networks* (pp. 13–18).

Cha, M., Mislove, A. & Gummadi, K. P. (2009). A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th international conference on world wide web* (pp. 721–730).

Cheng, J., Adamic, L., Dow, P. A., Kleinberg, J. M. & Leskovec, J. (2014). Can cascades be predicted? In *Proceedings of the 23rd international conference on world wide web* (pp. 925–936).

Choi, D., Han, J., Chung, T., Ahn, Y.-Y., Chun, B.-G. & Kwon, T. T. (2015). Characterizing conversation patterns in reddit: From the perspectives of content properties and user participation behaviors. In *Proceedings of the 2015 acm on conference on online social networks* (pp. 233–243).

Domingos, P. & Richardson, M. (2001). Mining the network value of customers. In *Proceedings of the seventh acm sigkdd international conference on knowledge discovery and data mining* (pp. 57–66).

Dow, P. A., Adamic, L. A. & Friggeri, A. (2013). The anatomy of large facebook cascades. In *Proceedings of the 7th international aaai conference on weblogs and social media.*

Goldenberg, J., Libai, B. & Muller, E. (2001). Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters*, *12*(3), 211–223.

Gonçalves, P., Araújo, M., Benevenuto, F. & Cha, M. (2013). Comparing and

combining sentiment analysis methods. In *Proceedings of the first acm conference on online social networks* (pp. 27–38).

Gonzalez, R., Cuevas, R., Motamedi, R., Rejaie, R. & Cuevas, A. (2013). Google+ or google-?: dissecting the evolution of the new osn in its first year. In *Proceedings of the 22nd international conference on world wide web* (pp. 483–494).

Granovetter, M. (1978). Threshold models of collective behavior. *American journal of sociology*, 1420–1443.

Guy, I., Avraham, U., Carmel, D., Ur, S., Jacovi, M. & Ronen, I. (2013). Mining expertise and interests from social media. In *Proceedings of the 22nd international conference on world wide web* (pp. 515–526).

Hartigan, J. & Wong, M. (1979). Algorithm as 136: A k-means clustering algorithm. *Applied Statistics*, 100–108.

Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The elements of statistical learning*.

Hong, L. & Davison, B. D. (2010). Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics* (pp. 80–88).

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Machine learning: Ecml-98* (Vol. 1398, p. 137-142). Springer Berlin Heidelberg.

Jordan, A. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes.

Kairam, S., Brzozowski, M., Huffaker, D. & Chi, E. (2012). Talking in circles: selective sharing in google+. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1065–1074).

Kempe, D., Kleinberg, J. & Tardos, É. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the ninth acm sigkdd international conference on knowledge discovery and data mining* (pp. 137–146).

Koller, D. & Sahami, M. (1997). Hierarchically classifying documents using very few words. In *Proceedings of the fourteenth international conference on machine learning* (pp. 170–178). San Francisco, CA, USA.

Kouloumpis, E., Wilson, T. & Moore, J. (2011). *Twitter sentiment analysis: The good the bad and the omg!*

Kwak, H., Lee, C., Park, H. & Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on world wide web* (pp. 591–600).

Lerman, K. & Ghosh, R. (2010). Information contagion: An empirical study of the spread of news on digg and twitter social networks. *ICWSM*, *10*, 90–97.

Ottoni, R., Las Casas, D. B., Pesce, J. P., Meira Jr, W., Wilson, C., Mislove, A. & Almeida, V. (2014). Of pins and tweets: Investigating how users behave across image-and text-based social networks. In *Eighth international aaai conference on weblogs and social media* (pp. 386–395).

Paul, M. J. & Dredze, M. (2011). You are what you tweet: Analyzing twitter for public health. In *Icwsm* (pp. 265–272).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*.

Purver, M., Griffiths, T. L., Körding, K. P. & Tenenbaum, J. B. (2006). Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings of the 21st international conference on computational linguistics and the 44th*

annual meeting of the association for computational linguistics (pp. 17–24).

Ramage, D., Hall, D., Nallapati, R. & Manning, C. D. (2009). Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 conference on empirical methods in natural language processing: Volume 1* (pp. 248–256).

Rodrigues, T., Benevenuto, F., Cha, M., Gummadi, K. & Almeida, V. (2011). On word-of-mouth based discovery of the web. In *Proceedings of the 2011 acm sigcomm conference on internet measurement conference* (pp. 381–396).

Saito, K., Kimura, M., Ohara, K. & Motoda, H. (2016). Super mediator–a new centrality measure of node importance for information diffusion over social network. *Information Sciences*, *329*, 985–1000.

Sparck Jones, K. (1988). Document retrieval systems. In (pp. 132–142). London, UK, UK.

Sun, E., Rosenn, I., Marlow, C. & Lento, T. M. (2009). Gesundheit! modeling contagion through facebook news feed. In *Proceedings of the third international icwsm conference.*

Tumasjan, A., Sprenger, T. O., Sandner, P. G. & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, *10*, 178–185.

Wang, C. & Blei, D. M. (2011). Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th acm sigkdd international conference on knowledge discovery and data mining* (pp. 448–456).

Weng, L., Menczer, F. & Ahn, Y.-Y. (2013). Virality prediction and community structure in social networks. *Scientific reports*, *3*.

Yao, L., Mimno, D. & McCallum, A. (2009). Efficient methods for topic model

inference on streaming document collections. In *Proceedings of the 15th acm sigkdd international conference on knowledge discovery and data mining* (pp. 937–946).

Ye, S. & Wu, S. F. (2010). Measuring message propagation and social influence on twitter. com. In *International conference on social informatics* (pp. 216–231).

Yu, B. & Fei, H. (2009). Modeling social cascade in the flickr social network. In *Fuzzy systems and knowledge discovery, 2009. fskd'09. sixth international conference on* (Vol. 7, pp. 566–570).

Zhao, Z. & Mei, Q. (2013). Questions about questions: An empirical analysis of information needs on twitter. In *Proceedings of the 22nd international conference on world wide web* (pp. 1545–1556).