Analysis of Telegram, An Instant Messaging Service

Arash Dargahi Nobari
Faculty of Computer Science and Engineering Shahid Beheshti University:G.C
Tehran, Iran
a.dargahinobari@mail.sbu.ac.ir

Negar Reshadatmand
Faculty of Computer Science and Engineering Shahid Beheshti University:G.C
Tehran, Iran
n.reshadatmand@mail.sbu.ac.ir

Mahmood Neshati
Faculty of Computer Science and Engineering Shahid Beheshti University:G.C
Tehran, Iran
m_neshati@sbu.ac.ir

ABSTRACT

Telegram has become one of the most successful instant messaging services in recent years. In this paper, we developed a crawler to gather its public data. To the best of our knowledge, this paper is the first attempt to analyze the structural and topical aspects of messages published in Telegram instant messaging service using crawled data. We also extracted the mention graph and page rank of our data collection which indicates important differences between linking patterns of Telegram nodes and other usual networks. We also classified messages to detect advertisement and spam messages.

CCS CONCEPTS

• **Information systems → Social networks**; *Spam detection*; Content ranking;

KEYWORDS

Instant Messaging, Telegram, Spam Detection, Classification, PageRank

1 INTRODUCTION

Instant messaging (IM) services have changed the way people communicate to each other in recent years. Although, it is not new at means of communication but the emergence of smartphones and mobile broadband technologies stimulated the evolution of communication patterns between people and organizations.

Different aspects of instant messaging applications and their influence on communication have been studied in recent researches. For example, comparison of short messaging service (SMS) and WhatsApp IM [2], the usage pattern of snapchat [5], group communication patterns in WhatsApp [7], the user behavior analysis in mobile internet [9] and users behavior patterns in WeChat social messaging groups [6] have been studied in recent years. These researches are mostly based on interview and data survey methodologies and to the best of our knowledge, there is no study on member's activity of IM networks considering the topical and structural aspects of real data collected from IM networks.

CIKM'17 , November 6–10, 2017, Singapore, Singapore

ACM ISBN 978-1-4503-4918-5/17/11…$15.00
https://doi.org/10.1145/3132847.3133132

In this paper, we investigate the communication patterns in Telegram[1] IM service which has been one of the most successful IM services in recent years. According to the official Telegram blog[8], it has reached one billion messages per day only after 16 months of activity in 2014.

In this research, we gather the public data published in Telegram in order to analyze the structure of its network and the published content. There are three types of communication in Telegram which are one-one (user to user communication which is not in the scope of our study), one-many (known as a channel in Telegram) and many-many (known as a group in Telegram).

Moreover, Telegram provides the service delivery features using bot platform. This feature has been used in different domains such as Internet of Things (IoT) [3] and health care [4].

Popular channels and bots in Telegram have become a vigorous platform for mobile advertisement to introduce new products, services and etc to their audience. As a result, channel owners and bot developers try to increase their audience by providing useful content and services. In spite of Twitter, Facebook, and many other web based social networks, public content in Telegram is not searchable via search engines such as Google and Bing, hence, the owners of channels which are in their early stages, tend to advertise the name and address of their channels in other popular channels. As a result, a huge portion of messages in Telegram are devoted to channel advertisement. Thus, messages in Telegram can be classified into two classes which are channel advertisement (i.e. *spam messages*) and non-advertising (i.e. *ham messages*).

The main contributions of this paper include: First, design and development of a crawler to gather instant messaging data from Telegram, Second, generating and analyzing the mention graph between nodes of Telegram network. Third, We manually determined topics of all crawled channels and also labeled a subset of the crawled messages as *ham* or *spam*. Our dataset including all crawled data and tags is publicly available.[2] Finally, we used the state-of-the-art classification methods to detect spam messages on Telegram.

Our analysis indicates that there are major differences between one-many and many-many communication patterns in Telegram. In addition, the link structures between nodes of Telegram network (i.e. mention network) is extremely sparse and consists of several separated connected component which indicates the lower tendency of members to mention other members.

[1] https://telegram.org/
[2] https://github.com/arashdn/telegram-research

2 INTRODUCING TELEGRAM

Telegram is an IM service in which users can send text messages, photos, videos, stickers and files of any type. A message sender or receiver in Telegram can be a user, group or a channel. In addition to user-user messaging, channels and groups can be used to broadcast messages in Telegram as follows:

- **Groups:** A group in Telegram consists of a set of users who are interested in the same topic. All members of a group can send and receive messages in the group. Users can be invited by another member of the group or they can join it by a unique join link provided by the group administrators.
- **Channels:** Channels are a feature to broadcast public messages to a large number of users. A channel usually has one or just a few administrators who are the only one(s) can publish messages in the channel. An unlimited number of users can subscribe to each channel by channel username or its join link.

Despite Facebook, Twitter, and many other social networks, there is no friendship relation between users in Telegram. However, a user in Telegram has a private contact list to manage messaging with his/her, close friends. As a graph, the nodes of Telegram graph can be users, groups or channels with several types of relationship between them as follows:

- **Forwarding:** A user or channel can forward an original message (published by another user or channel) to a different user, group or channel. Message forwarding in Telegram is similar to email forwarding and the *retweet* action in Twitter. The content of a forwarded message and the original message are the same but the forwarded message includes the name and profile link of the original content provider.
- **Mentioning:** Users and channels in Telegram can have a unique username which can be used as a reference to mention a channel or user in a message. Each message may contain zero, one or several mentions of usernames which start with @ character.

2.1 Crawling Telegram Data

We developed a crawler to gather different types of Telegram nodes (i.e. users, groups, and channels) and their relationships (i.e. mentions and forwards). We selected 185 public channels and groups as seed nodes. The crawler fetches the messages published by seed nodes and extracts the associated mentions, join links and forwards to explore new nodes in Telegram graph. The seed nodes were selected by 20 people with a diverse set of interests including general news, sports, education, entertainment, computer science, free talk, etc. Crawling process continues by crawling posts of these nodes as well as exploring new nodes which are mentioned in posts or there is a forwarded post of them. It is worth mentioning that all of the crawled data including nodes and the relationships between them are public data visible to each member of the Telegram.

3 GENERAL STATISTICS OF THE CRAWLED DATA

We crawled the data of Telegram for two weeks starting from 1-Oct-2016 until 14-Oct-2016. Table 1 indicates general statistics of the crawled data in this period. As shown in the table, more than 75% of explored nodes by the crawler are *users*, 23.9% are channels and only 0.5% are groups. Interestingly, more than 83% of the crawled content is published in groups.

The reason of existing very small portion of groups versus users and channels is that groups are usually used for private group messaging in Telegram and therefore, our crawler was able to find only a few number of public groups. Actually, Telegram groups are chat rooms with many-many conversation pattern which constitute more than 83% of all messages in our crawled data. The channels and groups in Telegram are different in many aspects. While channels are usually used to broadcast messages, groups are used to do conversations. Specifically, the rate of message publishing in channels and groups is different. The average rate of message publishing in groups and channels are 281.5 and 4.07 message per day, respectively in our crawled data.

As indicated in Table 1, channels are mentioned 31,914 times in crawled messages while users are mentioned 1,945 times. As we will explain in Section 4, the mentions in Telegram are used for different purposes like channel advertising, link exchanging, self-mentioning, content referencing and etc. Our data collection indicates that channel advertising and link exchanging, account for a larger share of mentioning than the content referencing.

As explained in Section 2, forwarding is an important relationship between nodes of Telegram. In a forward relationship, there are three participant nodes which are a) the original message provider, b) the message forwarder and c) the message receiver. In Telegram, the message receiver can be a user, a group or a channel, but the original message provider and the message forwarder can be a user or a channel. The related statistics are reported in Table 1. A forwarded message to a specific user is only visible for that user and crawler can not access them.

4 ADVERTISEMENT DETECTION

The mentioning relationship between channels in Telegram described in Section 2 can form a graph structure which is demonstrated in Figure 1. In this graph, each node represents a channel and each edge denoted as $u \rightarrow v$ represents a message published in channel u which mentions the name of channel v.

Three types of mention edges can be detected in our data.

- **Self mention**: The channel administrators sometimes mention the name of their channels at the end of each published message. Therefore, when their messages are forwarded or copied, then the name of the original channel is shown in the content of the forwarded message. The self-loops in the graph indicate this type of mentions.
- **Spam mention**: This type of mention happens when channel u publishes a message only to introduce a set of channels by mentioning their names. In this type of mention, the message does not include actual content. This type of mentions are indicated by a red edge in Figure 1.
- **Ham mention**: A message including actual content published in channel u, which mentions channel v, indicates ham mentions which represented by gray edges in the graph.

As indicated in Figure 1, A considerable portion of edges are spam mentions which do not have any actual content. This kind of

Table 1: General statistics of the data collection crawled from Telegram. The - sign for user column indicates private data which the crawler cannot access them.

	Channel	Group	Users
Count of nodes	2,556	54	8,080
Ratio of nodes	23.9%	0.5%	75.6%
Count of nodes with at least one crwaled post	755	54	-
Count of posts	36,928	182,425	-
Ratio of posts	16.8%	83.2%	-
Count of mention	31,914	-	1,945
Count of original posts published by	7,907	-	7,978
Count of posts forwarded by	12,962	2,915	-

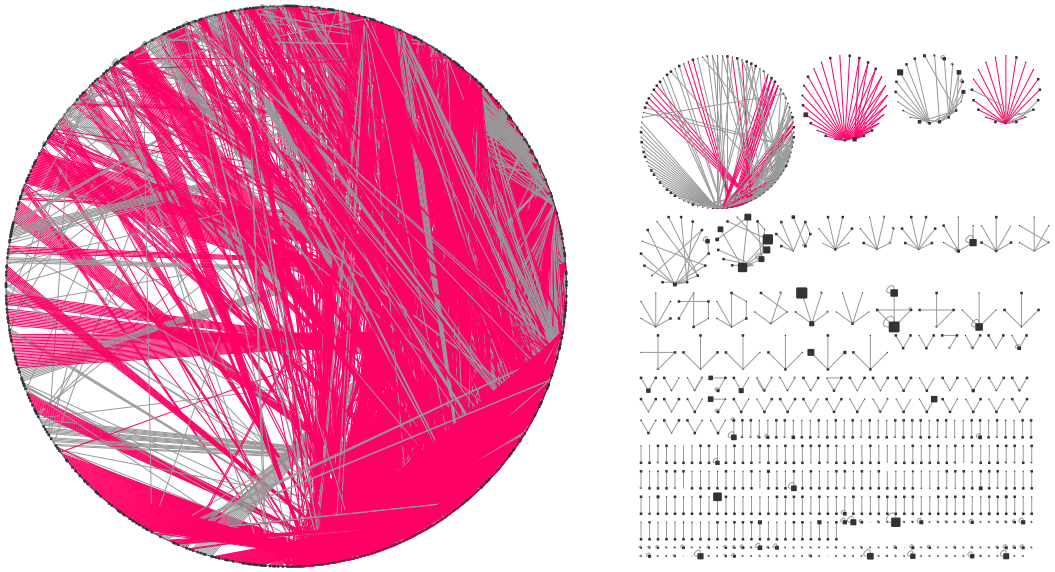

Figure 1: Mention graph of collected data

messages are mainly link exchanges and channel advertising. Detection and removing these messages is a prerequisite for analyzing the actual content of Telegram. The state-of-the-art classification methods can be used to discriminate spam messages. We manually labeled a randomly chosen subset of our data collection including 5270 messages. Three people independently labeled each message as spam or ham and majority voting approach is used to determine the label of each message.

In this short paper, we used neural network, SVM and decision tree classification methods trained using features indicated in Table 2. The results of these algorithms are summarized in Table 3. The overall performance of classifiers are almost the same. Our experiments indicated that features f_2 and f_7 are the most important features to distinguish ham and spam messages. Specifically, if the number of mentions in a message is more than five, It can be a strong signal to detect spam messages. In addition, forwarded messages from other channels are more probable to be spam messages. We used decision tree classifier to label all messages. Then, spam messages are indicated by red color and ham messages by gray color.

Table 2: Features to detect advertisement messages

	Feature	Type
f_1	Message Length	Numeric
f_2	Number of @ in Message	Numeric
f_3	Number of links in message	Numeric
f_4	Number of Telegram links in message	Numeric
f_5	Time of sending Message(Hour)	Nominal
f_6	Time of sending Message(Minutes)	Nominal
f_7	Is a forwarded Message	Boolean

Table 3: Comparison of various methods for advertisement detection

Method	Percision	Recall	F-Measure	Accuracy
Neural Network	0.857	0.818	0.837	0.805
SVM	0.820	0.834	0.827	0.799
Decision Tree	0.861	0.807	0.833	0.798

5 ANALYZING AND DISCUSSION

In this section, we describe our main findings and observations of mention graph.

(1) **The structure of graph**: The mention graph indicated in Figure 1 includes 2059 nodes (i.e. channels which mentioned another channel or were mentioned by another one) connected to each other by 18,296 mention links. 89% of edges are spam mentions, 8% are ham mentions and 3% are self mentions. The size of nodes in the graph indicates number of followers of the channel (i.e. nodes). Intuitively the number of followers of the channel can be considered as an indicator of the channel popularity and its quality. Surprisingly, In spite of web graph[1], there is no relationship between the degree of nodes in Telegram mention graph and its number of followers. Specifically, many popular channels in Telegram are standalone nodes or they have a low degree. On the other hand, a lot of nodes in the largest connected component of mention graph i.e. high-degree nodes, are not popular channels.

(2) **PageRank vs. Number of channel followers**: PageRank is a way of measuring the relative importance of nodes in graph[1]. In order to compare the PageRank of nodes in mention graph and their importance (i.e. number of their followers), we discretized number of followers of a channel into 10 groups and the PageRank level into 7 groups using equal frequency binning. Equal frequency binning is a binning method that ensures that the bins contain approximately the same number of records. The heat-map indicated in Figure 2 visualizes the relationship between the PageRank score and the popularity of channels. The darkness of each cell indicates the number of associated channels in that cell. A surprising result here is that there are lots of channels with a huge number of followers having the lowest level PageRank. This observation indicates that the traditional PageRank method can not be used to detect high quality channels in Telegram.

(3) **Topical Linking**: In order to enhance analysis of the linking behavior of channels in Telegram, we first defined 15 topical categories and then manually assigned each channel to the most relevant topic. Figure 3 indicates the topical linking pattern observed in our data. An interesting observation here is that topic t_2 (Entertainment) and t_3 (Advertisement) have many mention links with almost all other topics except t_2, t_3, t_8 and t_{14} which most of them are related to NEWS agencies and they usually have a website. Thus they can attract members without participating in link exchanges, while other topics(e.g. general, education, art and etc) usually do not have a website, hence, they rely on link exchange to find new members.

6 FUTURE WORKS

Analyzing the messages and mention graph in Telegram, indicates that the PageRank Algorithm will not fit for this network. Therefore, An appropriate algorithm should be proposed to rank Telegram channels. In addition, more advanced classification methods can be proposed to detect spam messages.

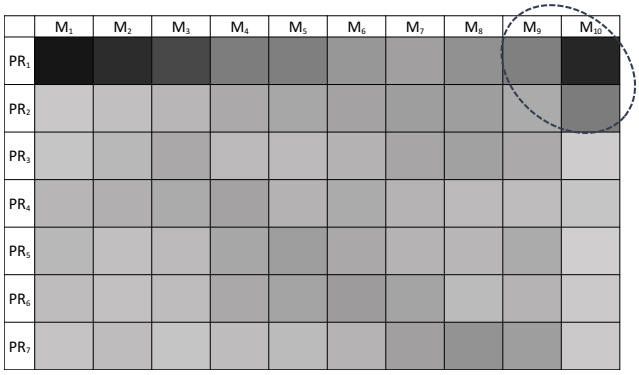

Figure 2: Heat-map for PageRank and number of followers bins

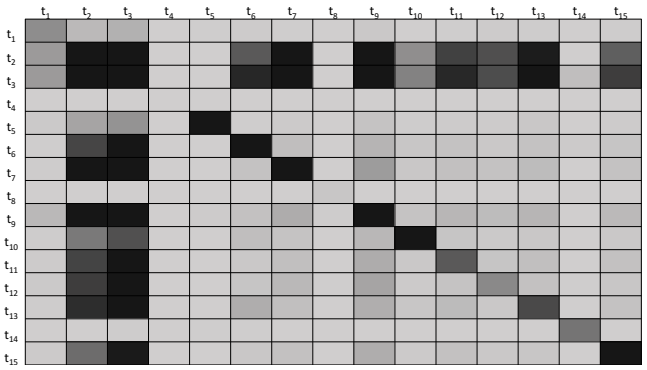

Figure 3: Heat-map for topical relation of mentions

REFERENCES

[1] Sergey Brin and Lawrence Page. 1998. The Anatomy of a Large-scale Hypertextual Web Search Engine. *Comput. Netw. ISDN Syst.* 30, 1-7 (April 1998), 107–117.

[2] Karen Church and Rodrigo de Oliveira. 2013. What's Up with Whatsapp?: Comparing Mobile Instant Messaging Behaviors with Traditional SMS. In *Proceedings of the 15th International Conference on Human-computer Interaction with Mobile Devices and Services (MobileHCI '13)*. 352–361.

[3] J. C. de Oliveira, D. H. Santos, and M. P. Neto. 2016. Chatting with Arduino platform through Telegram Bot. In *2016 IEEE International Symposium on Consumer Electronics (ISCE)*. 131–132.

[4] Mohtasham Ghaffari, Sakineh Rakhshanderou, Yadollah Mehrabi, and Afsoon Tizvir. 2017. Using Social Network of TELEGRAM for Education on Continued Breastfeeding and Complementary Feeding of Children among Mothers: a Successful Experience from Iran. *International Journal of Pediatrics* 5, 7 (2017), 5275–5286.

[5] Lukasz Piwek and Adam Joinson. 2016. What do they snapchat about? Patterns of use in time-limited instant messaging service. *Computers in Human Behavior* 54 (2016), 358 – 367.

[6] Jiezhong Qiu, Yixuan Li, Jie Tang, Zheng Lu, Hao Ye, Bo Chen, Qiang Yang, and John E Hopcroft. 2016. The lifecycle and cascade of wechat social messaging groups. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 311–320.

[7] Michael Seufert, Anika Schwind, Tobias Hoßfeld, and Phuoc Tran-Gia. 2015. *Analysis of Group-Based Communication in WhatsApp*. Springer International Publishing, 225–238.

[8] Telegram. 2014. Telegram Reaches 1 Billion Daily Messages. https://telegram.org/blog/billion. (2014). (accessed May 2, 2017).

[9] J. Yang, Y. Qiao, X. Zhang, H. He, F. Liu, and G. Cheng. 2015. Characterizing User Behavior in Mobile Internet. *IEEE Transactions on Emerging Topics in Computing* 3, 1 (2015), 95–106.