Capstone Project Description - M7

Alexandre Sablayrolles
sla382@nyu.edu

Christina Bogdan
ceb545@nyu.edu

Varun DN
vdn207@nyu.edu

Abstract

In studying social media data, network perspectives are valuable in terms of understanding the connections across different types of entities, and for characterizing the connections and entities themselves. The purpose of this project would be to develop a tool to visualize these networks, which in some cases can contain thousands of nodes and hundreds of thousands of edges. This tool should provide a quantitative way to display these networks with as little information loss as possible, with the ability to embed further information and annotations, in a reproducible way. The challenge is computational (e.g. applying forcedirected layout algorithms) but also analytical (part of the solution could lie on network reduction techniques, such as community detection algorithms, or backbone decomposition).

1 Data

In the project description, the dataset that we would be using is described as "retweet/mention Twitter networks collected by the Social Media and Political Participation lab at NYU. The datasets related to worldwide political events, such as protests in Turkey and Egypt, political discussion in the US, etc." Ideally, our dataset would include actual tweet text alongside the representation of the network. Because we are not sure yet if this data is available to us or if contains all the information that we need, we consider multiple ways to collect Twitter social network data.

1.1 SMaPP Lab

The dataset described in the project outline was collected by the SMaPP lab at NYU (1). We will be meeting with Duncan Penfold-Brown, who proposed our project, in the coming week to see if the original data is still available to us. It is our understanding that the data is already processed to represent a network, and is not just raw tweets.

Given that this already processed dataset is not available, the SMaPP lab provides some tools for working with twitter data and it might be possible to recreate the proposed dataset (2)

1.2 Twitter API

We also have the option to collect data directly from the Twitter API using the Python library Tweepy (3). This would require more pre-processing before the data is in the right format to visualize a

network. Twitter's API does have limitations on the data that one can collect - we would not be able to collect historical streaming data and there would be limits on the number of tweets we would have access to. On the other hand, having the raw data would give us a lot of flexibility. Indeed, it makes it possible to analyze the dynamics of the network (who twitted who and when, who started to follow who, etc.)

1.3 Historical Tweet Data

Our advisor, Lauro Lins, mentioned that he had already been collecting geo-tagged tweet data in the US from Twitter's API since February and that we could use this dataset. This could be useful if we want to look at trends that have been evolving over the past few months, such as the US election.

1.4 Scale

Twitter data, and more generally Web-scale social media data, is very large. Dumping data from Twitter feed takes approximately 30 gibabytes per day. Also, Twitter has a lot of users in the US (dozens of millions). This scale represents a new challenge if we do not get the curated data from SMaPP. In this case, we will apply massive preprocessing to focus on a particular topic (e.g. US election), that will reduce the size of the data. Also, we are considering clustering users into representants to get a clearer representations. This will lead to a more computationally effective approach, as well as a clearer visualisation.

2 Proposal

Our end goal would be to create an interactive tool to visualize large-scale network of tweets. We aim to use Python for data preparation, and D3.js to create our visualization. The objective will involve visualizing these networks in a compact manner without losing much information. Different visualization techniques and theory from social network analysis will be used/implemented within the tool. Below we outline some popular network visualization tools that already exist, and then outline some ideas that we had for our visualization.

2.1 Related Work

1. **Gource**

   Gource (4) is an OpenGL-based 3D visualisation tool for source control repositories. The repository is displayed as a tree where the root of the repository is the centre, directories are branches and files are leaves. Contributors to the source code appear and disappear as they contribute to specific files and directories.

   The idea can be similarly extended to visualize tweets about a certain topic or person.

2. **Gephi**

   Gephi (5) is an open-source software for network visualization and analysis. It helps data analysts to intuitively reveal patterns and trends, highlight outliers and tells stories with their data. It uses a 3D render engine to display large graphs in real-time and to speed up the exploration.

2.2 Directions

2.2.1 Gephi adapted to political analysis

The original project as described in the Capstone class features a lot of common points with Gephi. However Gephi might not be suited for political analysis nor have the exact set of features that are needed. Thus, a close collaboration with the SMaPP lab could help us define a roadmap that is both innovative and useful for them.

2.2.2 Message-passing visualisation

Twitter dynamics are bursty, as was originally described in (6). This paper describes the two-way interaction between message passing over a social graph, and changes in the graph (i.e. creation/deletion of social links). The authors show that the social structure is steady over time, except when there is massive retweeting, and this leads to a lot of changes in the social structure.

This phenomenon is very interesting, and we think that a proper visualisation of it would lead to deeper insights. Our *a priori* idea of such a visualisation would be to show the two dynamics (message passing and social changes) on the same graph, but computed independently. The (x, y) coordinates of a node would correspond to the social structure (i.e. the node is close to its followers). To show retweets and how the message ripples over the network, we will show the starting node getting bigger and smaller, and then its followers getting the same expansion/explosion bubble display. The idea is that the message is a wave that propagates to nearby nodes.

In this regard, factorizing users into representants would lead to a stronger signal in terms of retweet, and hopefully eliminate some of the noise.

@inproceedingsML14, author = Myers, Seth A. and Leskovec, Jure, title = The Bursty Dynamics of the Twitter Information Network, booktitle = Proceedings of the 23rd International Conference on World Wide Web, year = 2014,

References

[1] Social Media and Political Participation Lab
`https://wp.nyu.edu/smapp/`

[2] Social Media and Political Participation Lab. Data Collection and Analysis Tools.
`https://wp.nyu.edu/smapp/data-collection-and-analysis-tools/`

[3] Tweepy - An easy-to-use Python library for accessing the Twitter API.
`http://www.tweepy.org/`

[4] Gource - software version control visualization
`http://gource.io/`

[5] Gephi - The Open Graph Viz. Platform
`https://gephi.org/`

[6] Myers, Seth A. and Leskovec, Jure, The Bursty Dynamics of the Twitter Information Network, *Proceedings of the 23rd International Conference on World Wide Web, 2014*