Twitter Project Notes

The project will utilize a high-performance computing (HPC) resources including a local small cluster containing 2 nodes, 10 cores, 25 Gigabytes of memory and resources provided by the Janus Linux Cluster. The project will set aside 10 Terabytes of local storage for project activities. In addition, the project will provide archiving utilize Google Code hosting service to provide long-term archiving and future collaboration. The service (<http://code.google.com/projecthosting/>), Wordpress local web server. Also, the service provides project wikis to allow for project management. Project and data management services include:

* a permanent URL
* secure replicated online storage location (multiple copies of data)
* accurate metadata
* a globally accessible repository
* the option for contextual linking between data, and published research results
* RethinkDB Access

The project will examine each research question through, experimental analysis of data streams. The social media sites Google+, Twitter, and Facebook all provide live data streams: In which actors on such sites public data can be programmatically extract using free/existing application programming interfaces (APIs) to provide a never-ending stream of data.

**RQ1:** How does socio-technological interactions impact trust within the physical world based on cultural elements with information generated from an academic library?

**Data:** The data streams will enable the project to conduct linkage-based, structural analysis to determine communities, links, and an overview of global evolution behavior with respect to different aspects of engagement.

**RQ2:** How do ubiquitous social media facilitate or hinder patterns of temporal trust between individuals, declared groups or public entities?

**Data:** The data streams mentioned above provide a rich set of temporal characteristics of social media based engagement in which evolving communities develop and disappear. All of the social media services that will be utilized in this project provide mobile information within the data stream information such as Twitter which offers GEO information. This service provides information on users location, vicinity to events. This information will enable us to measure dynamic preferential attachment, in which links between actors can be assessed on their heterogeneous preferences for other actors.

**RQ3:** How do individuals or groups form dynamic communities within social media, which include trust?

**Data:** The data streams will provide the project with the necessary data (followers, friends, keywords, language, gender, and geo-location), to enable the implementation of algorithms to infer linkage, which are not yet know in the social network world, thereby enable the prediction of trust associations within the online communities.

* **RQ4:** To what extent do network structures emerge based on library engagement elements in both unique and non-unique ways within particular social media platform such as Twitter?
* **Data:** In social media networks nodes in the network maybe labeled using tags, these attributes can provide information about the propagation of information, in which information is correlated to network structure. Therefore, tagging information provides significant information about influence across actors, who are the most influential actors for influence spread i.e. trust among the group
* **RQ5:** How will user names and ids be analyzed for social engagement elements with community-generated content? Quantify
* **RQ6**: How the set of streaming APIs offered by Twitter give developer low latency access to Twitter's global stream of Tweet data.
* **RQ7:** How we store some information on the server that will remain even after the server shuts down, and then we’d like to display this information to the user.
* **RQ8:** How can observing twitter data be used to determine the popularity and semantic of a specific book, and how would doing so be useful in library management.

**RQ9:** How can the observation of a social network lead to the discovery of the most important node?

**Data:** The term "most important" is not as straightforward as it appears; importance may have several different aspects such as degree centrality, Eigenvector Centrality, and betweenness centrality. Together, these aspects are known as "centrality measures". To illustrate the centrality aspects we will monitor the retweet network and the amount of followers user. Using this obtained data from the social network we will be able to clarify the most important person. Node.js will be used to extract the needed information.

**RQ10:** With all the vast amounts of tweets being exchanged among users how do you find the topic that you desire within the text?

**Data:** The latent Dirichlet allocation (LDA) is a topic modeling method that can be used to gather the most probable words for a topic. Each topic located in the LDA is a grouping of words. Every word contains a specific probability of being related to topic, but the probability the word receives is different depending on the topic. LDA retrieves words with the highest percentage for a given topic. We will use LDA topic modeling with node.js.

**RQ11:** How can we build a lexicon automatically using data collected directly from Tweet data that contain emoticons as labeled data?

**Data:** We will download a lexicon directly from npmjs.org. Data collected from twitter does not give specific positive or negative labels. Any data collected from Tweets that contain emoticons such as :) or :( will be labeled as positive or negative.

**RQ12:** How can the stream data of a social network be analyzed using sentiment based on a sentiment score that is derived from captured tweets?

**Data:** The sentiment analysis affiliates text with a sentiment score. A sentiment score is either positive or negative, and a sentiment analysis reflects the emotion of the user's piece of text. Accumulating the sentiments will help us to illustrate how people are reacting to a specific organization, product, event, or a topic.  **Data Analytics with Visualization**

**RQ13:** What graphical representation would be adequate in displaying the most important person in the network based on retweets and followers.

**Data:** In previous research questions we discussed about centrality where we can find important people and concepts in the network. We will use information Flow Networks and Friend-Follower networks to gain insight into how and why users interact. An important factor in a network is to find out the original source of information and the intermediate users along the information propagation path. A highly used method for visualizing the network is the *force-directed layout*. We will use this implementation in D3.

**RQ14:** How can force-directed layout be used for visualizing network information and retweet propagation ?

**Data:** In network visualization nodes need to have adequate spacing between them, but nodes that are connected to each other should be placed as close as possible to each other. The Force-Directed Layout solves this problem while at the same time keeping the results visually easy to understand. For example, if there was a natural disaster like an earthquake, a first responders would be more interested in tweets that contained reports of damage or power outages. This is an example of applying the Force-Directed Layout to a retweet network. To accomplish this we create groups of words called topics. To explain, we generate topics with #damage and #poweroutage; we would then color code both topics for visualization purposes. Before can view the visualization, we must extract and configure the topics. We do this operating the method *ConvertTweetsToDiffusionPath* in class *CreateD3Network.* Visualizing the network will be used with the methodcreate\_network which can show us the size of the node which indicates its importance in the network. The size of the node increases with more retweets. The nodes will be color coded according to their topic preferences. Also, the graphic will show us the larger nodes and smaller nodes of the network with the producer as well as the consumer of the information. Likewise, the Friend-Follower visualization can be visualized using the same methods.

**RQ15:** What methods and tools would be appropriate to view time sensitive temporal information that analyze information relating with time and present a natural ordering of time-oriented information?

**Data:** Time-series information is usually visualized with one x-axis representing time, and they also include a y-axis that represents another dimension like the amount of Tweets and the velocity between tweets and retweets. We will use a method known as the *GenerateDataTrend* defined in the class *ExtractDatasetTrend.* Zoom and filter are two operations which can be use to make investigating data easier. Brushing & linking and focus+context are two approaches that can improve your experience and add to the visualization discussed above. TrendLine.js will be used for Brushing & linking and focus+context. In addition with using Time-series information, we can determine which hashtags are more popular based on the volume of tweets for multiple topics.

**RQ16**: **What is control chart (Reminder)**

**RQ17:** How can maps be used to visualize location information, and how can the maps be used to effectively be used to analysis geographic data?

**Data:** One way to effectively visualize location information is to use tweet location. Tweet location can easily be used to highlight single tweet locations. The individual tweets can be marked as dots on a graph; each of the dots are known as markers. We are able to customize the shape, color, and style of each marker to correspond to operation requirements. *OpenStreetMaps* uses tiles to present different colored dots. Tiles are rendered images on a map. Sometimes the dot method is not a viable option. This can be because there are too many dots, causing the data to be unreadable. We will use heat maps to represent the data on the maps. Using **Kernel Density Estimation** (**KDE**) we will be able to created a *heatmap* from the Tweet locations. KDE will be used to generate estimates from the Tweets.

**RQ18:** How can textual visualization determine the most popular or discussed topic in a network?

**Data:** Word Clouds can be generated to use important words in text. The amount of times a word is used determines its frequency. Frequency can be an important tool when determining the importance of a word. Word Cloud displays the importance of a word with the size of its font. We will need to sacrifice punctuation and correct grammar to gain additional characters. Words must first be broken down into tokens to generate a word cloud. The frequency of the tokens will be determined by using the *GetTopKeywords* method. Word clouds come with disadvantages though. The job of understanding what the words mean is placed on the reader. However, we can get passed this problem by taking the top keywords and organizing them into broad topics. A topic chart can represent the time and volume of usage of individual words. We can create this chart by using the *CreateTopicChart.* The dimensions are not static and can be changed with other dimensions with something like location.

**RQ19:** How can tweets be analyzed and visualized along with the four dimensions: temporal, geo-spatial, network, and textual?

Data: *TweetXplorer* can be used to evaluate temporal, geo-spatial, network, and textual information all at one time. This tool is capable of displaying multiple dimensions of data on the screen at one time.