Social Network Analytics—Final Project

# Overview:

“The ***Deepwater Horizon* oil spill** (also referred to as the **BP oil spill**, the **BP oil disaster**, the **Gulf of Mexico oil spill**, and the **Macondo blowout**) began on 20 April 2010 in the [Gulf of Mexico](http://en.wikipedia.org/wiki/Gulf_of_Mexico) on the [BP](http://en.wikipedia.org/wiki/BP)-operated [Macondo Prospect](http://en.wikipedia.org/wiki/Macondo_Prospect" \o "Macondo Prospect). It claimed eleven lives and is considered the largest accidental marine [oil spill](http://en.wikipedia.org/wiki/Oil_spill) in the history of the petroleum industry, an estimated 8% to 31% larger in volume than the previously largest, the [Ixtoc I oil spill](http://en.wikipedia.org/wiki/Ixtoc_I_oil_spill" \o "Ixtoc I oil spill). Following the [explosion and sinking of the *Deepwater Horizon* oil rig](http://en.wikipedia.org/wiki/Deepwater_Horizon_explosion), a sea-floor [oil gusher](http://en.wikipedia.org/wiki/Oil_gusher) flowed for 87 days, until it was capped on 15 July 2010. The US Government estimated the total discharge at 4.9 million barrels (210 million US gal; 780,000 m3). After several failed efforts to contain the flow, the well was declared sealed on 19 September 2010. Some reports indicate the well site continues to leak.” [<http://en.wikipedia.org/wiki/Deepwater_Horizon_oil_spill>]

While the oil spill originated from a leak that started in April 2010, the crisis unfolded over a period of several months (until the leak got capped) and its economical, ecological and political repercussions have been ongoing ever since. Not only did the spill affect the environment but also changed the way communication is conveyed during crisis. The oil spill was, according to statistics released by Twitter, the most microblogged about topic of 2010, which revealed wide interests and concerns in the “Twitterverse.” It triggered an outpour of reactions on the part of multiple stakeholders, including the public at large, local communities, news media, activist organizations, government, and British Petroleum (BP), the firm most directly associated with the spill ([Hoffman & Devereaux Jennings, 2011](#_ENREF_46); [Kirsch, 2010](#_ENREF_58); [Vaast et al., 2012](#_ENREF_98)).

You are hired by NetworkThink, a think tank specialized in studying the implications of social media and social networking websites on business and the society, to study the unfold of the Deepwater Horizon crisis on Twitter and to analyze the role and communication of various stakeholders involved in the event.

# Dataset:

NetworkThink has acquired a dataset containing 70,000 tweets related to the crisis during its first three months (April 2010-July 2010). The dataset was acquired through Topsy (http://topsy.com), a publicly available real-time search engine for the social web. Topsy archives tweets and allows users of the service to search for tweets within a defined time period. This dataset is provided to you as a SQLite database file in Moodle. The database has four tables:

1. TWEET: this is the main table containing the set of tweets including among other information the tweeted, the tweeter, the content, the date, the score of each tweet.
2. TWEETER: this table contains the classification of 675 tweeters whom researchers in NetworkThink identified into seven categories: social movement organizations (SOCMOV), media companies (MEDIA), celebrities (CELEB), corporations (CORP), government (GOV), and other (OTHER) who are twitter accounts for stakeholders who have an established identity outside of twitter. In addition, there is one important category GRASSRT which is designated for grassroots tweeters who are active accounts on Twitter that do not correspond to established stakeholders in real life.
3. HASHTAG: this table lists the hashtags (#) used in each tweet in TWEET.
4. MENTION: this table lists the mentions (@) used in each tweet in TWEET.

Take a look at the database and its content. If you are not familiar with SQLite you can use SQLiteStudio a free GUI tool to interact with SQLite databases [<http://sqlitestudio.pl/?act=download>].

In analyzing this dataset, one can think of different ways to slice it into networks. In particular you can build:

1. A mention network: nodes represent tweeters and edges represent mentions. This network represents the flow of communication among tweeters.
2. A hashtag network: nodes represent hashtags and an edge represents the fact that two hashtags co-occurred in the same tweet.
3. An affiliation network (tweeter-hashtags) that models the association between tweeters and the hashtags they use in their tweets. This network would represent the association between various stakeholders and their main concerns or positions in the crisis.

# Tasks:

## Create and analyze the mention network

1. Starting from the database create the mention network representing communication among tweeters over the first three months of the crisis. Indicate and clarify your choices in modeling the network (i.e. directionality, weight, and other attributes).
2. Analyze the structure of the network. Does the network exhibit a small world structure? Plot the degree distribution. Is there an evidence of a power-law distribution?
3. Who are the most influential tweeters from each group? Which group is more influential on aggregate?
4. Is communication more concentrated within or across groups?
5. Identify the communities in the network based on communication patterns. Do those communities correspond to the predefined groups?
6. Produce a network plot that visualizes the patterns identified above.

## Create and analyze the hashtag network

1. Create the hashtag network. As in the previous task, indicate clarify your choices.
2. What are the most frequent hashtags? Classify them in themes (i.e. environment, news, etc …).
3. Group the nodes into the identified themes. Plot the network in groups. What is an
4. Create two snapshots of this network, one representing the first 15 days of tweets and the second representing the last 15 days. Is there a shift in terms of hashtag usage? Identify a potential emerging trend.

## Create and analyze the affiliation network

1. Indicate your choices for nodes, edges and their attributes?
2. What insights can you learn by examining the affiliation network?
3. Visualize the affiliation network separating the two categories of nodes.

## Sentiment analysis

In this task you will analyze the content of the tweets and try to detect expressed sentiments (negative vs. positive). You will be using a simple analysis based on word polarity. AFINN is a list of English words rated for valence with an integer between minus five (negative) and plus five (positive) [<http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010>]. There is even a small script that uses the AFINN ratings to classify sentences [<https://gist.github.com/fnielsen/4183541>].

1. Compute the polarity of each tweet in the dataset.
2. Which stakeholders are expressing more negative sentiments?
3. What hashtags are associated with negative/positive sentiments?
4. Re-plot the mention network (1.f) color coding the edges depending on the sentiment expressed in the tweet? Is there a pattern to observe?
5. Overtime, does sentiment get more positive or negative? Is this trend consistent across groups?

## Implications

* 1. What did learn from examining the oilspill dataset?
  2. Who were the most influential actors in the network?
  3. Is Twitter a new medium for media companies to disseminate information? Or is it a platform for the masses to express their ideas?
  4. Is there an evidence for social movement organizations benefiting from Twitter to call for action?
  5. What political party was most active?
  6. Were political parties engaging in a debate online or just relaying their pre-existing frames?
  7. Bonus: any other important findings?

Provide a clear answer and outline the data you used to reach the conclusion.

**SOLUTION:**

**3:)** (i) Nodes are Tweeters and hashtags. Edge represents tweeters who used particular hashtags in their tweets.

32,943 cases where tweeters and hashtags matched were filtered from the tables TWEET,TWEETERS and HASHTAGS. There are in all 3,353 vertices. The data is then divided into groups based on type, with one group containing hashtags with other seven groups as given in the question.

Box Layout Algorithm - > Force Directed

Graph type -> Spiral

(ii) From the graph it is visible that people in group ‘GRASSRT’ tweet the highest number of hashtags followed by ’OTHER’, then ‘MEDIA’, then ’SOCMOV’ and so on.

**Group Edges:**

|  |  |  |
| --- | --- | --- |
| Group 1 | Group 2 | Edges |
| HASHTAG | GRASSRT | 19496 |
| HASHTAG | OTHER | 3765 |
| HASHTAG | MEDIA | 3306 |
| HASHTAG | SOCMOV | 5436 |
| HASHTAG | GOV | 596 |
| HASHTAG | CELEB | 132 |
| HASHTAG | CORP | 211 |

**#bp** has the highest degree and highest frequency with more people from group GRASSRT > OTHER>MEDIA>SOCMOV are using this hashtag in their tweets.

**Top hashtags :-**

|  |
| --- |
| **Top hashtags** |
| #bp |
| #gulf |
| #p2 |
| #tcot |
| #oil |
| #bpoilspill |
| #obama |
| #environment |
| #blacktide |
| #eco |
| #wildfire |
| #ocean |
| #green |
| #birds |
| #teaparty |

(iii) The images are attached in the zip file which depicts people from which groups are tweeting more about which hashtags.