

**ANLY482 Analytics Practicum AY2015/2016 Term 2**

**Social Network Analysis of SGAG**

**Project Proposal**

**5th February 2016**

**Supervised by:**

**Prof. KAM Tin Seong**

*Associate Professor of Information Systems*

**Prepared By:**

Wang Shyan Ann

Ng Tse Siong

Sherman Yong Chin Wei

Table of Contents

[1. Project Overview 3](#_Toc442468914)

[1.1 Project Sponsor and Background Information 3](#_Toc442468915)

[1.2 Our Motivation 3](#_Toc442468916)

[1.3 Objectives 3](#_Toc442468917)

[2. Data 5](#_Toc442468918)

[2.1 Datasets Provided by SGAG 5](#_Toc442468919)

[2.2 Crawling 5](#_Toc442468920)

[2.3 Merging Data 5](#_Toc442468921)

[2.4 Merging Data 6](#_Toc442468922)

[3. Methodology 6](#_Toc442468923)

[3.1 Overview 6](#_Toc442468924)

[3.2 Multivariable Regression on tweet content vs Google Trends 6](#_Toc442468925)

[3.3 Google Trends Correlation 7](#_Toc442468926)

[3.4 Targeted Content 8](#_Toc442468927)

[3.5 Exploration 9](#_Toc442468928)

[Method 9](#_Toc442468929)

[Objectives 9](#_Toc442468930)

[ Identify primary demographics of the audience to these prominent followers 9](#_Toc442468931)

[4. Project Plan 10](#_Toc442468932)

[4.1 Project Timeline 10](#_Toc442468933)

[4.1 Risks and Limitations 11](#_Toc442468934)

[4.2 Tools 11](#_Toc442468935)

[4.3 References 11](#_Toc442468936)

# Project Overview

### Project Sponsor and Background Information

SGAG is one of Singapore’s humour brands, and is a company aimed at delivering quality content that engages and entertains. SGAG’s content is crafted to cater to Singapore audiences, driven by its philosophy: “To make every Singaporean’s day a better day”. Founded in 2012, SGAG’s content distribution platforms have since grown to include popular social media platforms such as Facebook, Instagram, Twitter and YouTube, as well as mobile and web applications.

Thus far, SGAG has not devised a data-driven method to back their assumptions about their users. As their co-founder mentions, the content generated is done through "feel" (SGAG's general sentiments about each target customer segment). Their customer base has also been broken down into demographic and psychographic levels. Through greater segment targeting by content creation and channel targeting, SGAG aims to generate growth and use all 6 platforms to capture and drive user numbers up.

### Our Motivation

SGAG's value is to consistently deliver Singaporean-flavoured content that is both funny and strikes a local chord. SGAG’s content is crafted around and based the common daily undertakings of Singaporeans and important milestones and events. These content are then delivered daily to audiences.

Its co-founder is a SMU Alumni who started SGAG in class and he has grown the brand to the current state it is in now of having over a million daily unique users across all its platforms. We find that through this project, we can get to learn more about how this startup functions and at the same time, put our knowledge into practice.

### Objectives

The aim of the practicum is to provide deeper insight into SGAG's social network across its multiple platforms, namely Instagram and Twitter.

Our client is the co-founder of the company and he seeks insights that can spur growth in SGAG's follower numbers. Through our analysis and research, we aim to help discover what kind of users are on each platform; the key engagement leaders for each topic; and how wide is the reach of these individuals.

The final deliverables will aim to:

1. Visualise the social networks of SGAG
2. Identifying the user segments who engage SGAG's content, as filtered by topics and their reach

In order to achieve these objectives, the project will aim to explore some specific questions in order to find new insights within the social networks as well as validate the client's opinions about their users.

The approach to this would be to formulate a list of hypotheses that would be tested through the different stages of our analysis. A high level overview of each hypothesis is provided in the form of unanswered questions relating to content performance. Some of these questions that will be explored over the course of the project are:

* What are the types of users that are browsing SGAG?
* Who are SGAG's key influencers among its users and how can SGAG better target these key influencers?
* What topics are most relevant for which users?
* What is the role of seasonality and annual trends in the online readership pattern of SGAG's users?

# Data

### Datasets Provided by SGAG

Currently, SGAG only uses the insights provided on Facebook Page Insights and SocialBakers to gauge the reception of its posts, and much of the data that they have access to have not been analysed on a deeper level.

They have provided us with high-level social media metric data extracted from its social media platforms, namely Facebook, Twitter and Youtube. This gives us the following datasets that present a generic aggregated representation SGAG's followers:

* Unique visitors, by day and month
* Post level insights: Total Impressions, Reach, Feedback
* Engagement Insights: Likes, Viewed, Commented

This does not assist us directly in mapping out SGAG's social network, and we would have to crawl for more data using the respective platforms’ API.

### Crawling

Initially, we planned on mapping out the social networks for SGAG's main platforms: Facebook, Twitter and Instagram. However, due to the inaccessibility of user data that can be extracted from Facebook, we decided to focus on Twitter first since we are able to extract social network data much more easily.

By crawling the data through Twitter API, we were able to extract SGAG's Twitter social network data. This gives us the following information:

* Followed/ Following relationship represented by Edges
* Names of Twitter accounts associated with SGAG and their followers
* Interactions with SGAG's posts (Favourites, Retweets and Replies)

Due to Twitter and Instagram's API's querying limit, we will have to spend some time requesting for data. We have arranged to do this within 1 week.

After successfully crawling the data, we will load it up into Gephi and begin our visualisation.

Here is an example of an expected network visualisation for a social media platform.

### Merging Data

The Tweet ID provided by SGAG per tweet will be mapped with the crawled data above, and used to plot networks that link each tweet with retweets, replies, likes, etc.

The Twitter API provides easy importing of Twitter network data. The imported data will then be prepared and cleaned in the following ways through the merging of duplicate edges to reduce data noise, and grouping of nodes via a cluster algorithm. Metrics and graphs of the network will also be generated.

### Merging Data

One consideration is the ease of data retrieval after storing the SGAG network data extracted. As such, data storage in a relational database such as MySQL is preferred due to its support of various file format exports. Furthermore, data stored via this method can be easily manipulated and accessed for visualisations and further analysis through external software.

# Methodology

### 3.1 Overview

In the table below we outline the algorithms/techniques that we intend to execute for a particular objective.

|  |  |
| --- | --- |
| Objective | Analytical Approach |
| Network analysis via Degree centrality, Betweenness centrality | * Social Network Analysis * Cluster Analysis |
| Plan what to publish based on characterisation of audience | * Multivariable regression * Cross reference of google trends data and content of tweet |

### 3.2 Multivariable Regression on tweet content vs Google Trends

With reference to trending topics on a particular day of a tweet, multivariate regression will be performed to relate trending topics to the popularity of a tweet (retweet, likes, etc).

The key variables that we intend to explore are elaborated in the table below:

|  |  |
| --- | --- |
| Variable | Importance |
| Retweets | This measure shows how many times a particular tweet is being shared by followers. We think this is interesting because it highlights the willingness of an individual to share the tweet, increasing the probability that the tweet was interesting. |
| Url clicks | This measure shows how many times users actually click on the shortened link shared within a tweet. Given the succinct nature of a tweet, users who click on outgoing links are likely to find the tweet more interesting than other tweets, since clicking on the link would mean interrupting the "flowing" nature while reading the Twitter feed. |
| Likes | Compared to Url clicks and Retweets, this measure is the mildest, indicating that the user probably found the tweet interesting, but wasn't compelling enough to share. |
| Engagement Rate | A consolidated figure to illustrate how many people who see a particular tweet eventually interact with it (out of the total number of people who saw the tweet), in the following ways/forms:   * Link clicks * Favourites * Retweets * Replies * Embedded media clicks * Detail expands * Shared via email * Permalink clicks * User profile clicks * Follows |
| Tweet Text | Although the effectiveness of jokes can be tough to evaluate from a linguistics perspective, our initial approach would be cross referencing the hashtags used in the tweet with Google Trends data (Searches & Events) |

Giving a perspective on the important key variables that affects the popularity of a tweet will aid in the formulation of content that have higher penchant of being a popular tweet

### 3.3 Google Trends Correlation

While planning what content should be created, the team content team usually base it on gut feel, and usually the popular ones are accompanied when it is with regards to a big event in Singapore (eg. GE 2015). We want to incorporate the use of Google Trends to provide a more data-driven approach for SGAG in determining when and what content to publish. By analysing content consumption patterns of previous content around various types of events, future predictions can be forecast and appropriate content can be delivered at the appropriate timings.

### 3.4 Targeted Content

Content created at SGAG is tailored for Singaporeans, and revolve around the milestones commonly encountered at different ages. For example, the typical 18 year old male Singaporean faces the prospect of enlistment into Basic Military Training (BMT), and would experience a mixture of emotions. SGAG takes milestone events like these makes humorous content on it. Below are the targeted age groups for SGAG, with some of the associated commonly met milestones:

|  |  |
| --- | --- |
| **Age Group** | **Milestone Content Topics** |
| 18 - 21 | Male: National Service (Basic Military Training), Relationship issues  Female: Entry to University, Student Exchange Programme, Relationship issues, Social Night |
| 22 - 25 | Male: ORD (End of National Service), Entry to University, Relationship issues, Social Night  Female: Graduation from University, First Job, Colleagues |
| 26 - 34 | Male: Graduation from University, First Job, Colleagues  Female: Family, Having Kids |

Content creation is also based on events that happen in Singapore. These are categorized into 2 types, expected and unexpected. Expected events include mainstream events like the National Day Parade, while unexpected events include train breakdowns. A more comprehensive list is given below

|  |  |
| --- | --- |
| **Event type** | **Event Content Topics** |
| Expected | National Day Parade, SG50, SEA Games, Elections |
| Unexpected | Train breakdowns, different takes on Minister comments, Traffic accidents |

By understanding the content consumption habits of SGAG's social media audiences through further analysis, SGAG will be able to better craft content publishing strategies to increase consumer base.

### 3.5 Exploration

Following the following objectives, this is our framework to approaching the Twitter Network Analysis:

1. Identify most loyal followers of SGAG, based on their activity with SGAG’s content, i.e. Retweets, Favourites, Mentions
2. Identify network of SGAG, the demographics and networks of their most loyal followers
3. Breakdown segments of followers by interests, demographics, and other dimensions.

Methodology:

|  |  |
| --- | --- |
| **Method** | **Objectives** |
| Get 100 retweet IDs for each of all SGAG's posts, to find out who are the most loyal followers | * By obtaining the IDs per post, we narrow down to who are SGAG’s most loyal and active followers. * Also, identify the categories of tweets (topics) that segments of followers react most to |
| Get 2nd level IDs of the most active users who retweet, favorite and mention | * Users who on average retweet more than 3 different SGAG tweets[[1]](#footnote-1). This can show us the audience of these active followers |
| Generate list of followers with the largest follower base | * Identify opportunities for SGAG to collaborate and increase their reach to the followers of these followers with high-degree centrality.  Identify primary demographics of the audience to these prominent followers |
| Table Analysis | * Percentage of followers with default Twitter profile picture   + Characteristic of inactive or bot users * Number of new users   + Based on account creation date by year, e.g. 2015, 2016 * Average follower demographics:   + Average number of followers per user (including and excluding possible bots),   + Average number of friends per user,   + Average number of lifetime status posts per user (how active are their users) |

# Project Plan

### 4.1 Project Timeline

|  |  |
| --- | --- |
| **Week** | **Plan** |
| Week 2 | Researching on suitable areas for project exploration  Finding data sources |
| Week 3 | Exploring analytics tools (Gephi, NodeXL, etc)  Gathering data  Proposal submission |
| Week 4 | Going through analytics tools (Python, SQL, etc)  Crawling data from Twitter API |
| Week 5 | Crawling, cleaning & organising data  Classifying Tweets according to topics |
| Week 6 | Visualising network with Gephi  Finetuning visualization |
| Week 7 | Mid-term Presentation & Report |
| Week 8 | Generating analysis based on visualization |
| Week 9 | Analysing unexplored areas on completed network analysis |
| Week 10 – 11 | Recording findings and documentation |
| Week 12 | Documenting project’s limitations and areas for possible further research |
| Week 13 | Buffer |
| Week 14 | Final Presentation & Report |

### Risks and Limitations

|  |  |
| --- | --- |
| Risks | Mitigation Strategy |
| Lack of knowledge on data that can be obtained | Explore Twitter’s API  Consultations with the Supervisor to iron out problems and seek potential solutions |
| Data crawling takes too long | Employ different crawling algorithms through research and using available libraries |
| Inexperience with Analytical Tools | Explore the tools using smaller data sets to practice, and getting familiar with them before doing the actual analysis |

### Tools

Gephi, Python Library, MySQL

### References

Hogan, B. (2008). Analyzing social networks via the Internet. In N. Fielding, R. Lee, & G. Blank, The sage handbook of online research methods (pp. 141–160). Sage, Thousand Oaks, CA.

Hansen, Derek, Ben Shneiderman, and Marc A Smith. (2001). Analyzing Social Media Networks With NodeXL: Insights from a connected world. Morgan Kaufmann,

Visualizing Big Data: Social Network Analysis By Michael Lieberman (2014), Multivariate Solutions.

Charu C. Aggarwal. Social Network Data Analytics. Springer Publishing Company, Incorporated, 2011

1. Depending on the results, we will look for the average frequency of retweeting, and identify a frequency number that is around the 75th percentile of all active users. [↑](#footnote-ref-1)