Community Detection in Social Networks
Facebook and Twitter

A Thesis Proposal
Presented to
the Faculty of the College of Computer Studies
De La Salle University Manila

In Partial Fulfillment
of the Requirements for the Degree of
Bachelor of Science in Computer Science

by

FERNANDEZ, Ryan Austin
POBLETE, Clarisse Felicia M.
SAN PEDRO, Marc Dominic
TAN, Johansson E.

Charibeth K. CHENG
Adviser

July 14, 2016

Contents

Chapter 1

Research Description

This chapter is an overview of the research undertaken in the field of community detection in social networks. This chapter is divided into four sections which are the current state of the technology, research objectives, scope and limitations, and significance of the research.

1.1 Overview of the Current State of Technology

Social media has become much more prevalent in recent years. People can now participate in what is called microblogging, a way for people to share their thoughts, status, and opinions in short posts, like Twitter, where posts are limited to one-hundred and forty characters (Java, Song, Finin, & Tseng, 2007). As such, these social media platforms are a prime opportunity to mine sentiments and to detect communities in the social network.

Community detection is clustering multiple users into groups where users within the group are more similar than users from outside the group (Tang & Liu, 2010). Community detection is necessary because it observes the interaction of multiple users as opposed to common mining methods which may only deal with local predictions i.e. predictions based on a single node as opposed to predictions considering the entire network of users.

Numerous studies on community detection have already been done. Zhang, Wu, and Yang (2012) defined features; such as text content similarity, URL similarity, hashtag similarity, following similarity, and retweeting similarity; that can be used to identify similarity between two nodes and aggregated these similarities to detect communities. Their software just provided the listings of users in each community. Lim and Datta (2012) performed an inverted version in which they first defined interests and based on these interests, sought to extract communities from the network by identifying users that follow the top six celebrities, users with more than 10,000 followers, relating to the given interest. Their software only outputted the size and clustering coefficient of each community

Individual opinions of one specific user towards another was also studied by West, Paskov, Leskovec, and Potts (2014) combining sentiment analysis, using an L2-regularized logistic-regression classifier, with network analysis, inferring one users opinion of another by analyzing their common links with other users. Their output was the area under the curve of the receiver operating characteristic (AUC/ROC) and the precision-recall curves (AUC/negPR). In addition to these works, some visualizations have already been created such as SocialHelix by Cao, Lu, Lin, Wang, and Wen (2015), which uses the temporal extent of social communities, topics or events discussed, and user responses to topics and events to classify users into two sides of the argument. They then depict the two sides of an argument as strands in a double helix and their intersection defines events.

However, it is noticeable that all of these studies only involved Twitter and mainly used only common Twitter features for similarity analysis such as following networks, hashtag frequency, and retweet networks. There has yet to be a community detection tool that integrates data from Facebook, as well as the exclusive features from Facebook such as group membership and event membership, into the computation.

1.2 Research Objectives

1.2.1 General Objective

To produce a visualization of the detected communities on data found on Facebook and Twitter

1.2.2 Specific Objectives

1. To build a corpus of social media data;

2. To determine the various techniques and algorithms in detecting communities;

3. To determine the parameters/features to be used in detecting the communities;

4. To determine how to evaluate the correctness of the detected communities

1.3 Scope and Limitations of the Research

In order to perform a study on community detection, it is necessary to build a corpus of social media data. This research will cover searching for API's that will allow extraction of data from Facebook and Twitter.

Different techniques have been used in community detection. Among these techniques are the Infomap algorithm and the speaker-listener label propagation algorithm. This research will consider algorithms found in the review of related literature, including the Markov stability model, clique percolation method, k-means clustering, and divisive hierarchical clustering.

In the implementation of these community detection algorithms, it is necessary to identify which parameters/features/attributes indicate one users similarity to another. Inquiry will be done to identify these parameters, specifically how to extract them based on the raw data. The research will be limited to sentiment analysis and elements which can be extracted from a users post, which may include follow networks, hashtags, mentions, and retweets, which were mostly inspired from literature which focused on Twitter (Deitrick & Hu, 2013; Zhang et al., 2012; Lim & Datta, 2012). As such, Facebook specific features such as membership in groups and event participation may also be considered.

After community detection, it is necessary to determine whether the detected communities are sensible. Inquiry will be done to find appropriate metrics in determining the accuracy of detected communities. These algorithms will include average mutual following links per user per community or FPUPC (Zhang et al., 2012), modularity (Deitrick & Hu, 2013), and clustering coefficient (Lim & Datta, 2012).

1.4 Significance of the Research

Community detection is already a widely researched topic in the field of computer science. This research will contribute to that field by exploring a new domain, including Facebook in the scope of its community detection. This research can also contribute to the notion that community detection is a relevant field of study in this day and age.

This research can also be a very useful tool in the domains of viral marketing and political endorsement. This means that companies and governments may benefit from this research. Interested companies may use the result of this research to improve their sales and marketing. The government may use this to gauge public opinion on certain issues and to see analytics about which geographical areas have a particular opinion.

Chapter 2

Review of Related Literature

This chapter discusses the features, capabilities, and limitations of existing research, algorithms, or software that are related or similar to the proposed research. This chapter is divided into three sections. The first section discusses algorithms for community detection. The second section discusses algorithms for sentiment analysis and other similarity parameters used for community detection. The third section discusses community evaluation metrics.

2.1 Community Detection

Tang and Liu (2010) wrote a textbook on community detection. In the book, they sought to introduce characteristics of social media, review representative tasks of computing with social media, and illustrate associated challenges because with the emergence of social media websites, they felt it was an avenue to study human interaction and collaboration on an unparalleled scale. Multiple community detection approaches were discussed such as node-centric, network-centric, and hierarchy-centric.

Node-centric algorithms involve the maximum clique detection problem which involves searching for a maximum complete subgraph of nodes in a network graph that are all adjacent to each other. The clique percolation method can find overlapping communities by finding cliques of size k, and then producing a clique graph, wherein two cliques are connected if they share k-1 nodes. Each connected element in this clique graph is then a community.

Network-centric algorithms involves vertex similarity, which is the similarity of the nodes social circles or how many common friends the two nodes have. This is what structural equivalence deals with. Two nodes v_i and v_j are structurally equivalent if $\forall v_k \in \{x \mid x \neq v_i \wedge x \neq v_j\}\ e(v_i, v_k) \in E \iff e(v_j, v_k) \in E$, which is to say, v_i and v_j are connected to the exact same nodes. Nodes that are structurally equivalent belong to a community.

Hierarchy-centric algorithms come in two forms: divisive and agglomerative. In divisive clustering, the entire set of nodes starts out in one set and each time, each set is divided into two until each community only has one member. The division is done by finding the node with the lowest edge betweenness and removing it, since that node is most likely the node connecting two communities. Agglomerative clustering starts with each node in their own community and communities are joined if they increase the overall modularity of the set of communities. Modularity is given by

$$Q = \frac{1}{2m} \sum_{l=1}^{k} \sum_{i \in C_l, j \in C_l} (A_{ij} - \frac{d_i d_j}{2m}) \tag{2.1}$$

Where m is the number of edges, d_i is the degree of node v_i, C_l being the lth community, and $A_i j$ being the value in the adjacency matrix for node v_i and v_j.

These algorithms may be considered in the final community detection phase of the proposed research (Tang & Liu, 2010).

Lim and Datta (2012) aimed to detect communities that share common interests on Twitter, based on linkages among followers of celebrities representing an interest category because they wish to help markets identify target groups with common interests. However, their approach differs from the typical paradigm of "identify communities then, for each, identify interests", for they first identified interests they wish to extract communities from and from these interests, they then extracted the communities.

Given a set of celebrities, C, celebrities being users with more than ten thousand followers, the algorithm first gets the common followers of all the celebrities in the set using the formula.

$$P = \bigcap_{j \in C} (\bigcup_{i} link(i, c_j)) \tag{2.2}$$

Where $link(i, j)$ is given by

$$link(i,j) = \begin{cases} \{i\} & \text{i follows j} \\ \emptyset & \text{i does not follow j} \end{cases} \tag{2.3}$$

Given this set of fans, P, they used the Infomap Algorithm and the Clique Percolation Method to detect communities in P. For each interest they wish to extract communities for, they chose the top 6 most popular celebrities based on follower count. Google and Wikipedia were used to identify which interests a celebrity represents. Afterwards, all users that follow the 6 celebrities were selected. They then selected 200,858 random users as a control group. Their algorithm produced more communities and larger communities than the control group, as well as more consistent communities, having a higher clustering coefficient.

This paper provides an interesting alternative to detect communities by first specifying the interest of the community before detecting it, which may be used in the proposed research on top of the usual algorithms (Lim & Datta, 2012).

Zhang et al. (2012) sought to identify communities in Twitter based on common interests. This is because Twitter has become very popular recently but little is known of it in the user level. This study would help in user recommendation and tweet recommendation as well as viral marketing to specific target groups. To identify the communities, they first compute specific feature similarities, then aggregate these features to compute for the final user similarity, and then they used classical clustering algorithms to detect the communities. To identify the communities, they first compute specific feature similarities, then aggregate these features to compute for the final user similarity, and then they used classical clustering algorithms to detect the communities.

The specific features they used were textual contents. Each data point was the entirety of a users tweets. Latent Dirichlet Allocation was used to identify latent topics from the users tweets. URL similarity was also detected, finding which users share similar links. Hashtag similarity was also analyzed. The social structure of users was also analyzed, which includes following similarity and retweeting similarity.

In aggregating these similarities, the weighted sum of the previous similarities was computed to get the final similarity. Finally, k-means clustering was used to detect the communities based on their computed similarities.

This paper presents a possible framework for detecting the communities. The proposed research may even use similar features, in addition to the Facebook specific features, to detect similarity. The k-means clustering algorithm will be one of the proposed algorithms for use in the proposed research (Zhang et al., 2012).

Deitrick and Hu (2013) sought to use sentiment classification to analyze communities in Twitter because harvesting information from these online social networks (OSN) would aid in the fields of politics and marketing.

Their process is as follows: The follower network was represented as a weighted directed graph, each with initial weight of 1. To augment this, replies, mentions, retweets, hashtags, and sentiment classification of tweets were also harvested. These factors adjusted the weights in the graph. For community detection, the Infomap algorithm and Speaker-listener Label Propagation Algorithm(SLPA) were run.

Generally, the network with updated weights produced communities with greater modularity. Of the two algorithms, the Infomap algorithm performed better. Recurring sentiment analysis was also helped by performing the aforementioned algorithms on the accounts that have already been placed in detected communities, which permits more in-depth analysis into the users sentiment since it could be analyzed within the context of the detected community.

This research provides a new way to represent the network, with their updated weights, as well as more possible algorithms to consider in detecting communities (Deitrick & Hu, 2013).

Cao et al. (2015) proposed a visual analysis system, SocialHelix, because social media is a grand avenue for people to express their opinions and the researchers believed that an intuitive visualization that unfolds the process of sentiment divergence would have a far-reaching impact on multiple domains.

They first identified the key domain problems of social divergence before employing a data abstraction design to convert the raw data into a form that captures all the key factors of the aforementioned domain problems. This abstracted data is then represented in a visualization based on a visual DNA metaphor. In identifying the key domain problems, it is determined when divergences start and end, how they evolve, who is involved, what roles do they play, and why does divergence occur. In the abstraction phase, the raw data is decomposed into temporal extent of social communities, topics or events, and user responses to these topics or events. In the visualization phase, the opposite sides of the helix represent the two sides of a divergence. The helix curves represents the changes in the communities sentiment. Nucleobase pairs represents events that connect the two communities.

In implementation, the data was first filtered, removing unrelated posts and people. Statistical Linguistic Sentiment Analysis was used to determine the users sentiment. Finally, clustering was done using Hadoop, producing a cluster with 30 nodes.

In the end, all test users were impressed by the visualization and agreed with the researchers model for the visualization. All test users felt that divergence identification was made easy due to the visualization.

This research gives a sample visualization that can inspire the proponents own visualization, albeit the goal is not to model single divergences but an entire community. A possible tool, Hadoop, is also mentioned, which may be used to cluster the data (Cao et al., 2015).

2.2 Similarity Parameters

This section outlines the basis/features/parameters used in detecting communities. It is divided into two subsections. The first subsection deals solely with sentiment analysis. The second subsection deals with other network and node parameters not related to sentiment analysis.

2.2.1 Sentiment Analysis

Zhang et al. (2012) provided a formula to determine similarity in terms of text.

$$sim_{text}(i,j) = \frac{1}{\sqrt{D_{js}(i,j)}} \tag{2.4}$$

D_{js} is the Jensen-Shannon Divergence between the two users topic probability distribution given by

$$D_{js}(i,j) = \frac{D_{kl}(UT_i \mid\mid M) + D_{kl}(UT_j \mid\mid M)}{2} \tag{2.5}$$

Where $M = \frac{UT_i + UT_j}{2}$ and $D_{kl}(P \mid\mid Q) = \sum_{i \in topics} P(i) \log \frac{P(i)}{Q(i)}$ and UT_i is the probability distribution of user i for all topics. $UT_i[t]$ is the probability distribution for user i on topic t.

This research provides a metric to determine the similarity of two users in terms of post content, which can be used in the proposed research (Zhang et al., 2012).

Deitrick and Hu (2013) used a subjective/objective and positive/negative Naive Bayes classifier. To do this, all tweets were converted to lowercase; hashtags, usernames, urls were replaced with twitterhashtag, twitterusername, and twitterurl respectively; the tweet text was tokenized; repeated punctuation was replaced with the + sign e.g. !!! -¿ !+; sentence punctuation was split into separate tokens; non-sentence punctuation was removed. Ten-fold cross validation was used in training the classifier. Weights in the graph mentioned in section 2.1 were then updated if two users posted something with a similar sentiment and similar hashtag.

This research shows a clearly defined process in performing sentiment analysis, particularly the data cleaning step. This process could be adapted for the proposed research (Deitrick & Hu, 2013).

2.2.2 Other Parameters

Zhang et al. (2012) provided a few formulas to determine similarity in terms of URL, hashtag, following, and retweeting similarity.

URL similarity is given by the same formula as text similarity in section 2.2.1, only using links instead of topics.

Hashtag similarity is given by

$$sim_{hashtag}(i,j) = \sum_{k=1}^{n}(1 - \left|\frac{N_{ik}}{|H_i|} - \frac{N_{jk}}{|H_j|}\right|)\frac{N_{ik} + N_{jk}}{|H_i| + |H_k|} \tag{2.6}$$

Where N_{ik} is the number of times user v_i used the hashtag k while H_i is the total hashtags used by v_i.

Following similarity is given by

$$sim_{follow}(i,j) = \frac{c_{friend}}{\sqrt{|Friend_i|\,|Friend_j|}} + \frac{c_{follower}}{\sqrt{|Follower_i|\,|Follower_j|}} \tag{2.7}$$

$|Friend_i|$ is the total number of users v_i follows. $|Follower_i|$ is the total number of users that follow v_i. c_{friend} represents the two users common friends. $c_{follower}$ represents the two users common followers.

Retweeting similarity is given by

$$sim_{retweet}(i,j) = \frac{c_{retweet}}{\sqrt{|R_i|\,|R_j|}} + \frac{n_{ij} + n_{ji}}{|R_i|\,|R_j|} \tag{2.8}$$

R_i is the number of users whom v_i retweet. $c_{retweet}$ is the number of users both v_i and v_j retweet. n_{ij} is the number of times v_i retweeted v_j and n_{ji} is the inverse case.

The aggregate similarity is now given by

$$sim(i,j) = \gamma_t sim_{text} + \gamma_u sim_{url} + \gamma_h sim_{hashtag} + \gamma_f sim_{follow} + \gamma_r sim_{retweet} \tag{2.9}$$

With $\gamma_{feature}$ determined in a process described in section 2.3.

This research gives formulas that can be used in the proposed research to measure similarity (Zhang et al., 2012).

2.3 Community Evaluation Metrics

Zhang et al. (2012) used the average number of mutual following links per user per community(FPUPC) to evaluate their communities. Based on this, appropriate weights for the aggregation were found by first performing their k-means clustering algorithm using only one feature similarity for each of the similarities and extracting the FPUPC. Afterwards, they gave each feature similarity a weight based on the following formula.

$$w_{feature} = \frac{FPUPC_{feature}}{\sum_{f \in features} FPUPC_f} \tag{2.10}$$

The number of clusters, k, used in the k-means clustering algorithm was also tweaked to get the maximum FPUPC. They concluded that they were successful in generating relatively accurate communities due to the incrementally increasing FPUPC after adjusting the weights.

This provides one possible evaluation metric for the proposed research, as well as a method to provide weights for the feature similarities that the proponents will eventually be using for community detection (Zhang et al., 2012).

Table 2.1: Summary of Review of Related Literature

Reference	Community Detection Algorithms	Sentiment Analysis Model	Other parameters	Community Evaluation
(Tang & Liu, 2010)	Clique percolation method, similarity detection, divisive and agglomerative clustering			
(Lim & Datta, 2012)	Topic driven community detection, Infomap method, Clique percolation method			
(Zhang et al., 2012)	k-means clustering	Similarity Formula for Text	Similarity Formula for URL, Hashtag, Follower, and Retweeting	FPUPC metric
(Deitrick & Hu, 2013)	Weighted directed graph, Infomap Algorithm, SLPA	Subjective/Objective, Positive/Negative Naive Bayes Classifier	replies, mentions, retweets, hashtags	
(Cao et al., 2015)	Data abstraction design, Hadoop tool	Temporal extent of posts, topics and events, user responses, Statistical Linguistic Sentiment Analysis		

References

Cao, N., Lu, L., Lin, Y.-R., Wang, F., & Wen, Z. (2015). Socialhelix: visual analysis of sentiment divergence in social media. *Journal of Visualization*, *18*(2), 221–235. doi: 10.1007/s12650-014-0246-x

Deitrick, W., & Hu, W. (2013). Mutually enhancing community detection and sentiment analysis on twitter networks.

Java, A., Song, X., Finin, T., & Tseng, B. (2007, August). Why We Twitter: Understanding Microblogging Usage and Communities. In *Procedings of the joint 9th webkdd and 1st sna-kdd workshop 2007* (p. 56-65). Springer.

Lim, K., & Datta, A. (2012). Following the follower: Detecting communities with common interests on twitter. In *Proceedings of the 23rd ACM conference on hypertext and social media (ht12)* (Vol. 1, pp. 317–318). Association for Computing Machinery. doi: 10.1145/2309996.2310052

Tang, L., & Liu, H. (2010). *Community detection and mining in social media.* Morgan & Claypool. doi: 10.2200/S00298ED1V01Y201009DMK003

West, R., Paskov, H. S., Leskovec, J., & Potts, C. (2014). Exploiting social network structure for person-to-person sentiment analysis.

Zhang, Y., Wu, Y., & Yang, Q. (2012). Community discovery in twitter based on user interests. *Journal of Computational Information Systems*, *8*(3), 991–1000.