Topic-specified Campaign Design through Social Media

Jiajie Chen, Yue Ma, Chia-Jung Hsu, Yingjie Li, Xiaochen Zhang

I. Introduction - Motivation

As the social media continues to gain popularity and the innovating pace of its technology continues to accelerate, the potential business value of the online social media increases. One of the most promising application of today's social media is launching cost-effective campaigns. In this paper,
we propose a topic-specified campaign design algorithm that serves to optimize and facilitate the information diffusion through social media based on text-mining and human memory decaying theory. The proposed method identifies the specific targeted customers and takes advantage of the so-called "social media influencer" to reach them in a cost effective manner. Another pressing problem for the application of social media campaign is the lacking of a specific tool for campaign designers. As a result, we introduce a novel interactive visualization tool that can simulate the information diffusion process through time based on the configuration by the campaign designer. The simulation keeps track of the coverage and intensity of the campaign through time which enables the campaign designer to optimize the campaign length, frequency and so forth. Finally, we demonstrate our algorithm through Twitter data. The visualization can be useful to the campaign designers, business agents, sociologists, or even politicians and public policy makers. Thus, the objective of our study is to use the ubiquitous social media data to explore influencers that facilitate the information diffusion and create a visualization tool for the targeted users.

II. Problem definition

Currently, for campaign designers, there is no interactive visualization tool for them to simulate their campaign design. After our survey, the most discussed topics in social media researchers are network, message and user, all of which have been heavily studied in a separate manner. In our paper, we combine the existing ideas from all three areas and propose a multi-dimensional approach to filter our potential users and create an interactive interface to intuitive campaign design.

The major contribution of our study is two-folded. First, we introduce a comprehensive solution for a business campaign which combines the studies from above three elements: social network users, message, and network connectivity. For the specific topic provided by the social campaign designer, we will cluster the potential users and provide candidate influencers to be hired in the campaign. Second, we develop a campaign tool that enables flexible human-computer interaction through eminent visualization. Thus, we want to create one user interface which allows machine as an extension of human activity and human knowledge as a guideline for machine learning.

III. Survey

(i) Social Media Research Survey

According to reference [1], users, messages, and network characteristics are three major topics for social media researchers. Reference [2, 3] make a thorough review of indices to describe social media user behaviors and introduce new sets of attributes for social media user labeling, including natural of users, topics of interests and so forth. However, they only consider "seed" users, which ignores network diffusing beyond them. According to reference [4][5], current social media message analysis is performed through various text mining and natural language process algorithms, including topic modeling, clustering and sentiment analysis [5], which requires sufficient training data. As to network characteristics, reference [6] models the connectivity of the social media as a graph, most of the literature has been done to study the graph density, network distances and network cohesion [6].

However, there is a lack of effort to combine the information from all three aspects to achieve an optimal solution for business campaign. Moreover, there is no technology and study developed to provide a human-computer interactive channel which allows the merge of knowledge from both human and machine.

(ii) How does Twitter work?

Twitter is an online social networking tool in which users post 140-character updates of what is going on in their lives along with links to things they think are interesting, funny, or useful to their followers ("following" being essentially what "friending" is on other sites). People use twitter in many ways, some as a newsfeed by following prominent people or networks, some as a pseudo-chatroom by limiting their followers and whom they follow to close friends and family, and some as a microblog for updating people about the work they are doing and their personal lives.

IV. Proposed method

One of our contribution is proposing an optimized social media campaign design method which takes advantages of users, message, and network connectivity. The specific method consists of three major steps. The first step is identifying potential users and understand their user patterns based on the specified campaign topic. The second step is identifying the influencers that can effectively reach the potential users. The third step allows campaign designers to simulate the campaign flow in the social network for the given configuration before actually launching an expensive campaign.

(i) Clustering topics

We find potential topics from tweets using text mining. After we preprocess the tweets with term frequency inverse document frequency (TFIDF), we will get a tweet-words matrix. Then we aggregate the data by user and reduce the feature word dimension by selecting the top n key words. Next, we implement Non-Negative Matrix Factorization to find a lower-rank representation $W_{m\times k}$ and its lower-rank basis $W_{k\times n}$ of the original data matrix $X_{m\times n}$, where the objective function for the Non-Negative Matrix Factorization can be expressed as follows:

$$\arg_{W\geq 0, H\geq 0} \min \frac{1}{2}\|X - WH\|_{Fro}^2$$

Because of the non-negativity of W and H, we can interpret W as the user "soft" clustering results and interpret H as the weights of each feature word in the k clusters. This method is effective in real situation when we do not know numbers of topics in tweets. [8]

(ii) Identifying influencers and feature words for each topic

As we wish to find an optimal set of influencers to promote certain products, for example, cameras, we can use the clustering method to find the related user clusters (such as travel and photography) and identify their corresponding influencers. However, we need to point out our method assumes the campaign designer has already got an list of influencers so that we can recommend influencers to them intelligently base on their the preference of their followers.

In our dataset we already have a set of 20 influencers so we utilize the clustering result of each candidate to get the "preference" of each influencer. That will be used to estimate their scores on different topics.

It will be also useful for campaign designer understand what their potential customers are talking about. As we can interpret H as the weights of each feature word in the k clusters. We can sort out the top key words in each cluster according to their weights. This will be also useful for clustering evaluation. Since clustering is an unsupervised learning technique, the ground truth can be subjective, hence visually seeing the top

representative words of the clusters is critical to evaluate the quality of our clustering result.

(iii) Targeting audience

We identify potential audience of each influencer by defining active followers and their followers. The active followers are defined as followers who has interacted with influencers more than twice through retweets and comments. The followers of active followers are also defined using same measure. We define these followers as each influencers' audience because we assume our campaign information will diffuse from influencer to active followers and so on. Among these audience, we use the clustering result from previous section to define targeting audience in each topic. This is defined by calculating a threshold to represent level of interest in given topic. We can use mean, median, quartile, normalized "probability" or composite of these all to calculate the threshold.

(iv) User interface design

After we found the topics, we select top 5 - 10 topics to do our data visualization, depending on our needs. We're going to design an interactive website to provide a visualization tool helping users target influencers in topics they specify. The main information diffusion graph will be realized by D3- forced-directed graph with animated information flow to different users picked out using our model illustrated above.

V. Experiments/ Evaluation

(i) Clustering Topics & (ii) Identifying feature words for each topic

The nature of internet topics is long-tailed but topics are very centered at several. Figure 1 we use hashtag as a proxy of topics, and it is clear that in our dataset, several tens of hashtags have appeared for thousands of times. This guarantees a number of big clusters exists as a ground truth.

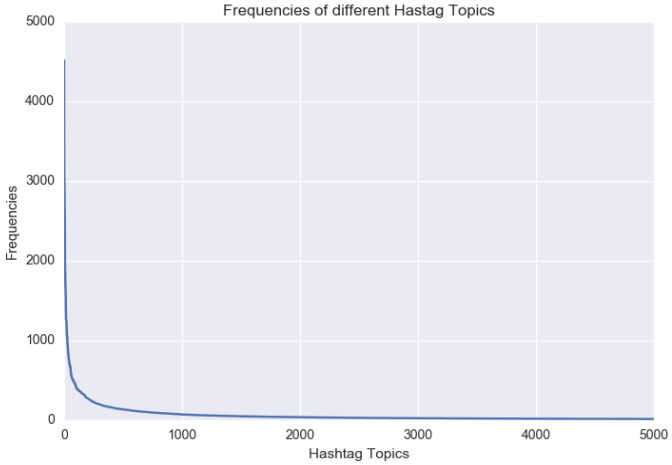


Figure 1.1: Hashtag Frequencies

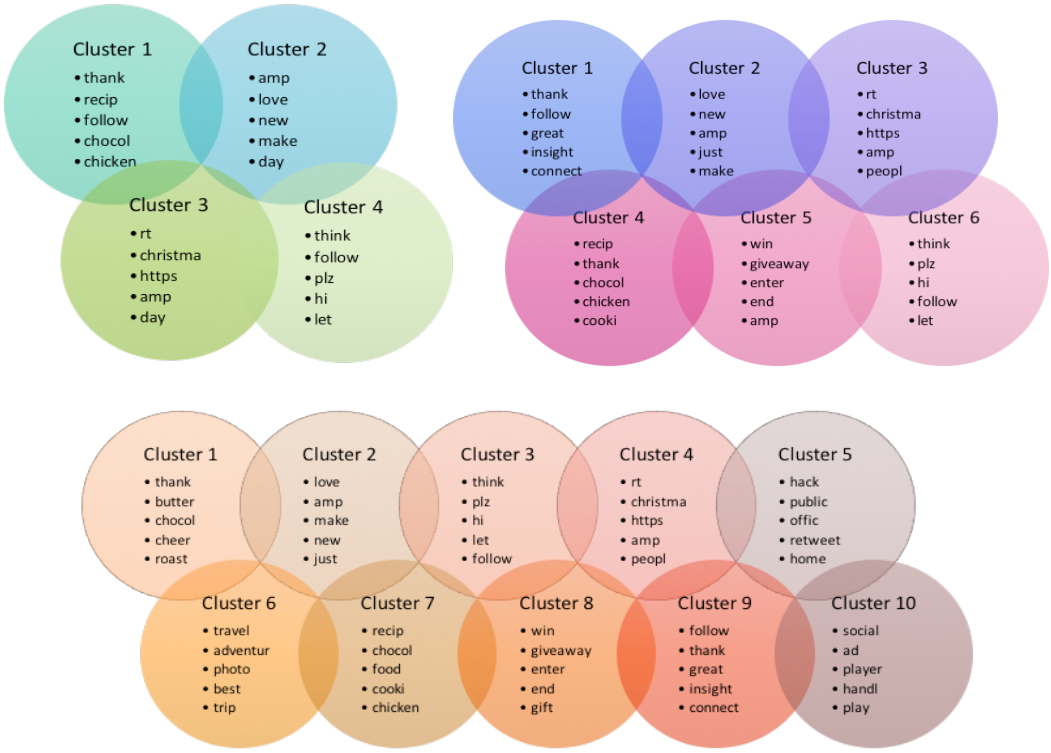


Figure 1.2: Experiments of 4 clusters, 6 clusters, 10 clusters

The goal of clustering is trying to minimizing within-cluster variance while maximizing between-cluster variance.

As Non-Negative Matrix Factorization clustered the data using matrix decomposition, the reconstruction errors are relatively high when number of clusters is small. This can be interpreted as information loss of is high when we try to abstract the original feature space with lower dimension. For example, when we abstract all human being using one word, it would be human, if we are allowed to use two words, then we are able to differentiate man and woman.

Figure 1.2 shows some results in our experiments. K=4 and K=6 failed to capture topics that differentiate users because users are clustered by positive words or greeting words such as thank, follow, love, etc. We can only identify a cluster for food when K=4 and K=6. However, when we choose K=10, the latent topics become more clear. It is easy to identify that cluster 1 and 7 are for food, cluster 6 is for travel. The other clusters also become easier to abstract if we look at more top key words in those clusters.

(iii) Targeting audience

in order to evaluate our targeted topic-specified audience, we summarized each user's hashtag usage frequency. by comparing users selected from top-related topic with their highly used hashtags, we validate those users like to topic-related hashtags as well, meanwhile, we could find other hashtags were being used. this could be potential business insight for agencies to find the right crowd of audience to promote their ads more effectively.

(iv) User interface Implementation

a. Topic Clustering

After our clustering analysis by, we found some keywords related in each of our clustered groups. We picked out 5 topics as a demo for our visualization, namely "Travel", "Food", "Holiday", "Design" and "Social Media".

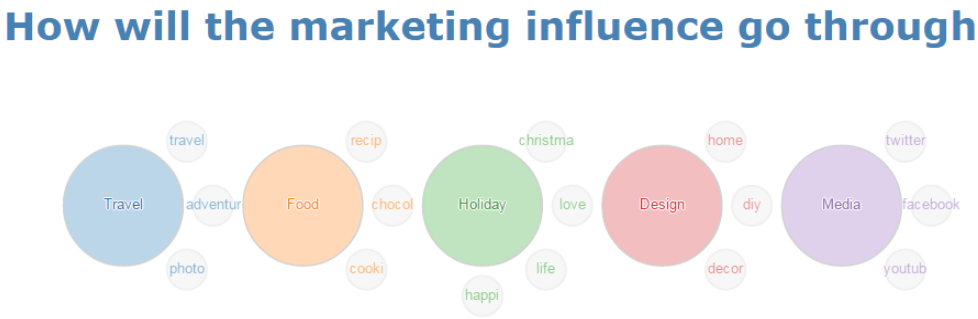


Figure 2.1: Topic

When the user's mouse points to one topic, the specific circle and its top related keywords will enlarge, where users can click on the keywords and jump to the relevant twitter webpage. In Figure 2.2, user points at "Holiday" topic, and the click of keyword "Christmas" will lead the user to the specific twitter page that highlighting "Christmas".

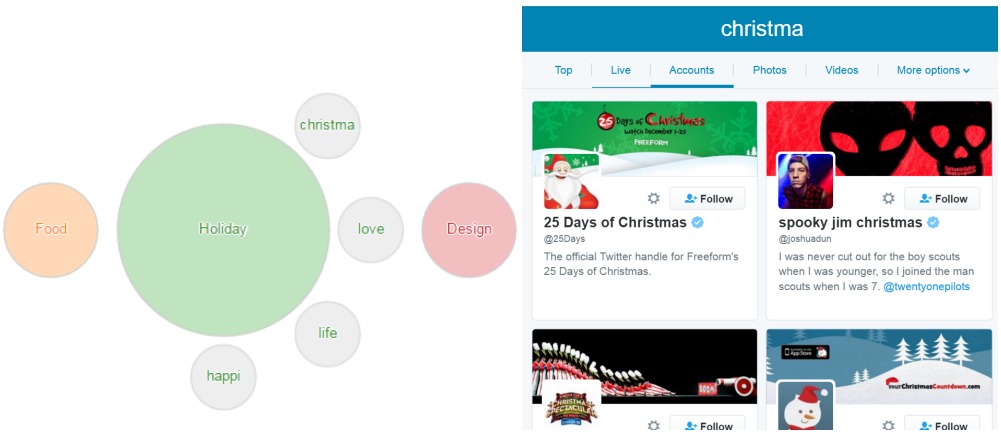


Figure 2.2: Mouse-over Animation and the Linked Twitter Webpage

Furthermore, we create the framework of the twitter network graph and simulation by using the test data. The test data is included the source node, target node and interactive time, which we use python to calculate number of each user's interactive edges.

b. Network Visualization

To help ad agencies visualize the information diffusion in Twitter network, this tool can let ad agencies select the inputs and see the network graph and simulation effect. We write a force graph visualization and trend simulation in d3.js.

We provide four inputs to select by users, such as "interested topic", "rounds of campaigns", "influencers numbers" and "the time interval between two campaigns". (Figure 3.1) Ad agencies can customize their campaign design by throw in different inputs.

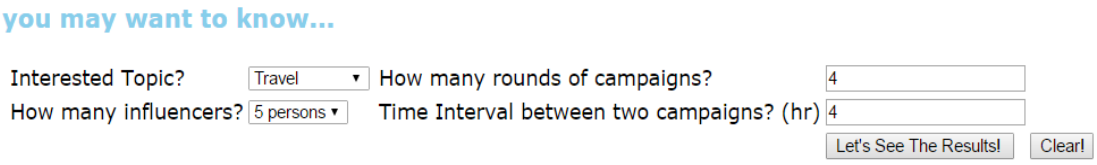


Figure 3.1 : User Interface Input Variables

After clicking the "Let's See the Results!" button, users can see the simulated network graph and the animation effect of the designed campaign.

As shown in Figure 3.2, nodes represent Twitter users and path represents following relationship. Basically we have 3 types of users: influencers (largest nodes), active followers of influencers (medium size nodes) and inactive followers of influencers (smallest nodes). What we need to notice is there are also some following relationships between two followers.

When there is no campaign, nodes show their default colors: influencer nodes are yellow (#31dfac), active follower nodes are light green (#82ebb7) and inactive follower nodes are light blue(#05a3a)

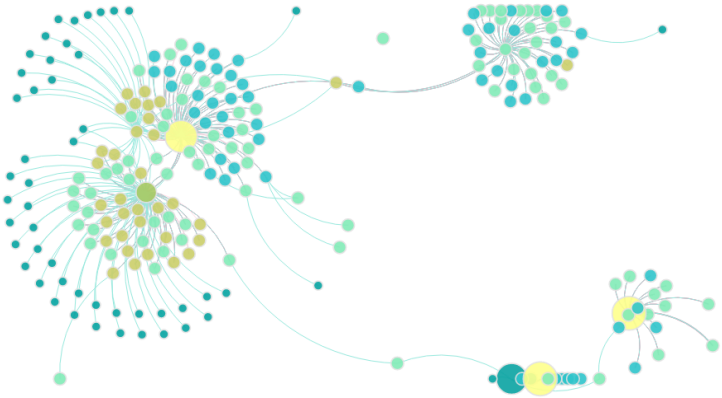


Figure 3.2 Default Look of Twitter User Network

After user of our interface specifies his/her interested topic, number of influencers, rounds of campaigns and time interval between two campaigns, he/she can click on "Let's See the Results!" and watch the animation for the campaign going through Twitter user network. In each round of campaign, there will be 3 phases:

In phase I, influencers post out the Tweet about a product, and can be understood as being "influenced" by their own post. The visualization effect is the influencer nodes are magnified and turning red (#db2059).

Figure 3.3: Figure Phase I

In phase II, followers of influencers will see the post. But we assume only those who are interested in this topic (evaluated by his/her score on this topic by our models) will be influenced and will help to spread out this information in ways including

retweet, creating their own Tweets on this, etc. The corresponding visualization effect is influenced followers' node being magnified and turn peach (#ff99cc).

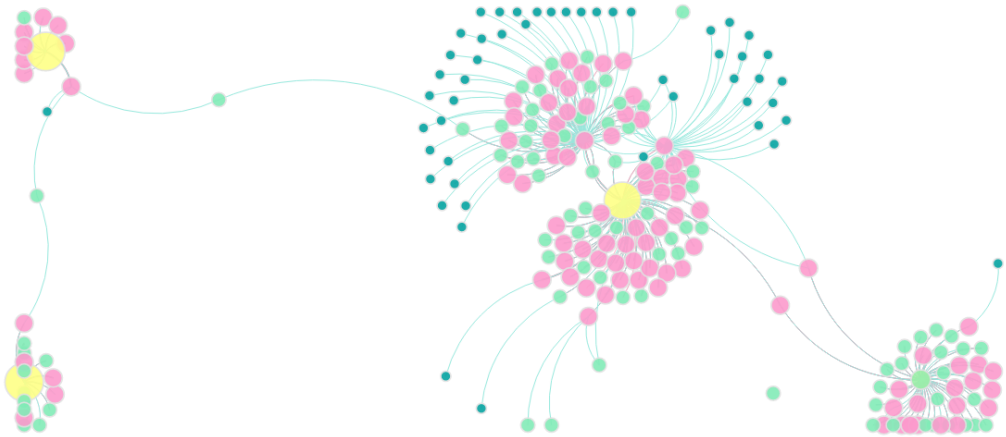


Figure 3.4: Figure Phase II

In phase III, follower's followers (found in present nodes) will see the post. We also assume only parts of them will influenced and the evaluation is their scores on this topic. They may also retweet or create Tweets on this topic, but we're not going to trace it due to the range of our research and limit of data. The animation of this phase is, followers' follower nodes will be magnified slightly and turn orange (#fbb30).

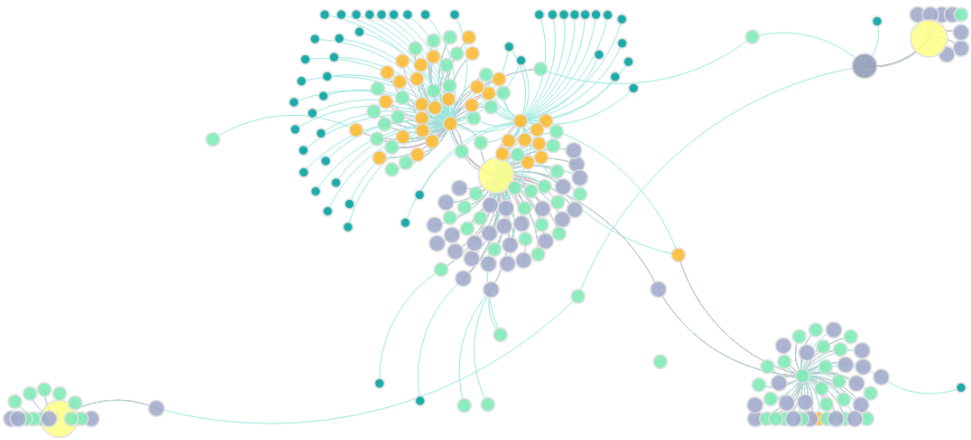


Figure 3.5: Figure Phase III

There are some other designs through this information diffusion process. First, we visualize the process of human memory through color changing of each node, where we assume most users will start to forget the campaign information after initial exposure. For this demo, we assume people will totally forget the campaign information in 3 hours, which is close to our general life experience. Therefore, you can observe that after being magnified and colored, nodes will gradually shrink in size and fade out to their original colors. Second, we assume the average time between two levels' user spreading the information is one hour. That is to say, after the influencers post out the ad, active followers may spread it out continuously, some within seconds,

some in hours, but the average time point will be 1 hour after the influencers' post. This is why you can see a time delay between different twitter users' node being magnified. It is worth mentioning that our simulation time scale is not real world scale: 2 seconds in simulation is equal to 1 hour in real world.

If the number of campaign user choose is larger than 1, our 3-phase animation will repeat for several times, as the user assigned.

c. Simulation statistics

While the network visualization provides campaign designers with a vivid and intuitive understanding about how their message is spreading in the social network, our simulation statistics quantify the effectiveness of the specific campaign. In Figure 3.2-4, we use animation to show the information diffusion through time, and we also show the number of users that are been exposed to the advertisement through time using an animated trend graph. The network graph and the trend graph are synchronized as shown in Figure 3.5.

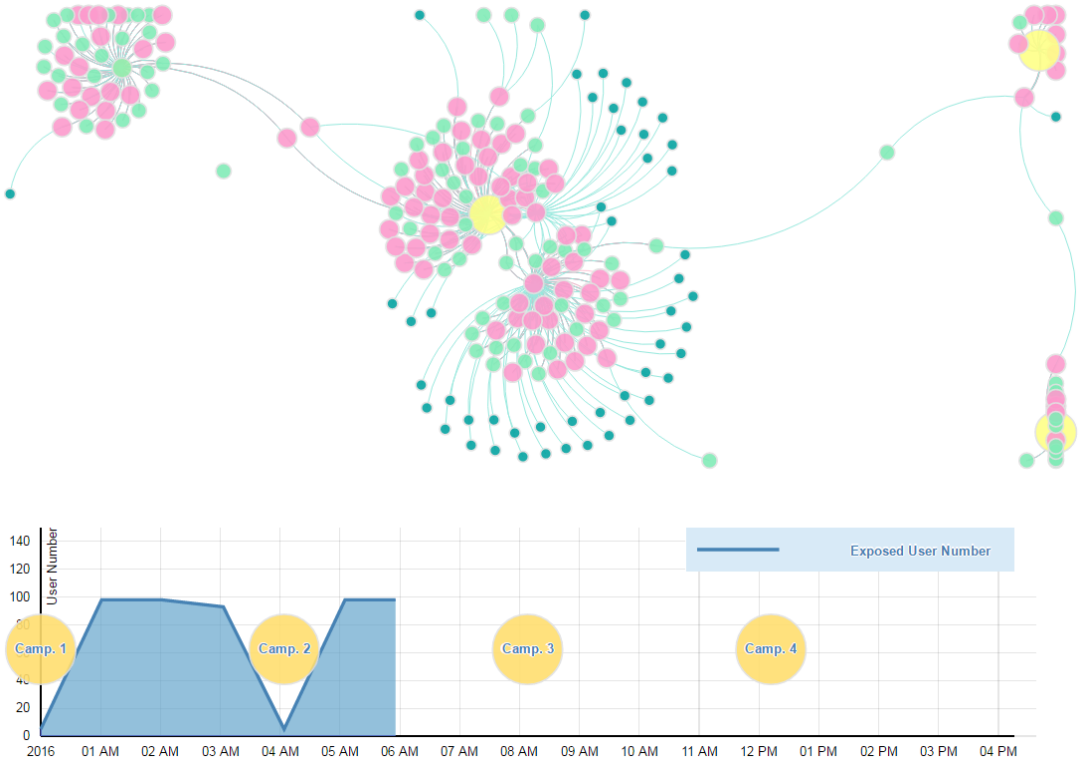


Figure 3.5 Network Chart and Trend Chart Synchronization

Another advantage of using the trend chart is allowing campaign designers to compare different campaign designs by changing the campaign setting. For example, if we choose to use 2 influencers in our initial design, we can see that during the pick time, we will have around 80 people being exposed to your campaign, as shown in Figure 3.6. However, if we want more people to get access to the campaign

information, we can hire more influencers and set the number of influencers to 5. Then, we will have an updated simulation statistic as shown in Figure 3.7, where we will have 100 people being exposed during the peak hours.

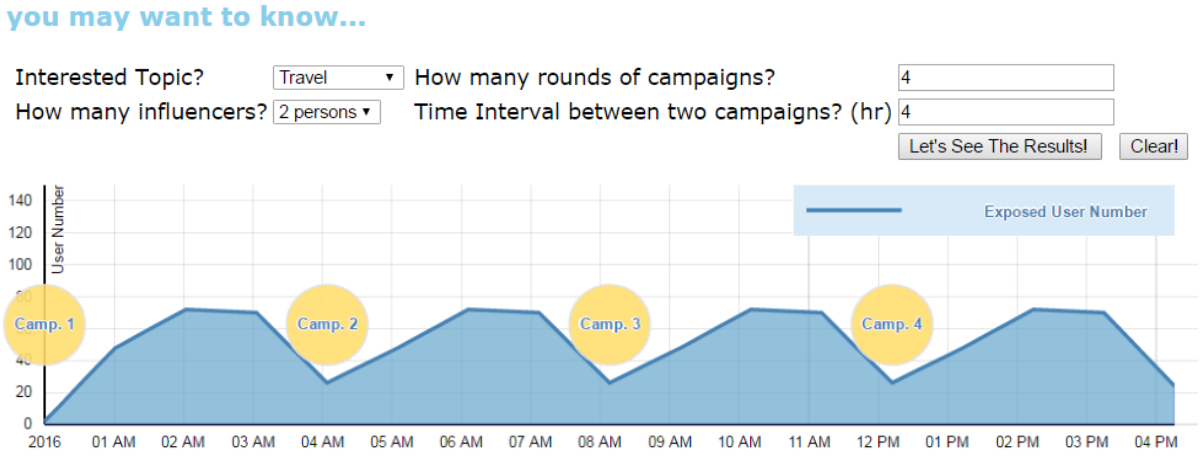


Figure 3.6 Campaign Design with Two Influencers.

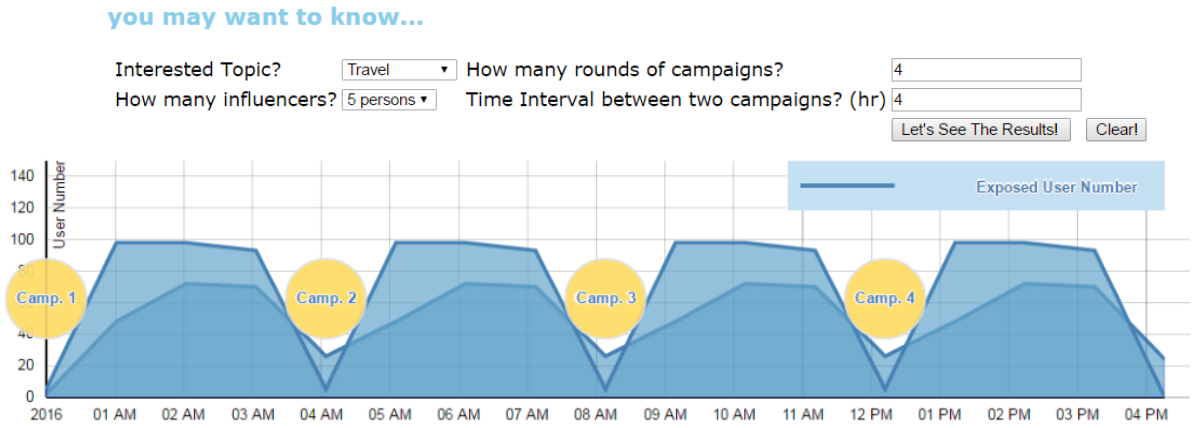
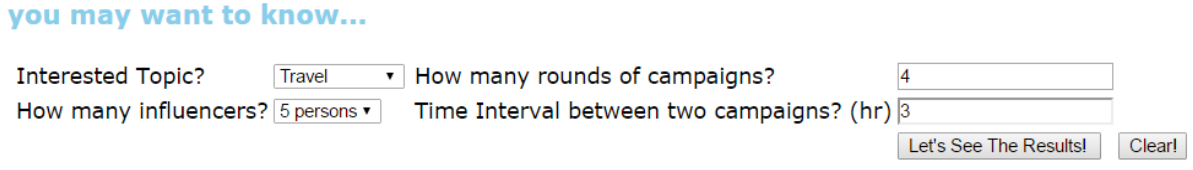


Figure 3.7 Campaign Design with 5 Influencers

Since our campaign design model also considers the human memory theory, we assume people will forget the campaign information 3 hours after the initial exposure. As a result, we can see from Figure 3.6 that, 3 hours after the campaign #1, the number of user being exposed to the campaign information will start to decrease until the second round of campaign kicks in. In fact, if we want to increase the intensity of our campaign, we can reduce the time interval between several social media campaigns. In Figure 3.8, we show that if we can reinforce user's impression on the campaign by reducing the campaign interval to 3 hours, where user will be exposed to the campaign information again and again before they memory fade away.



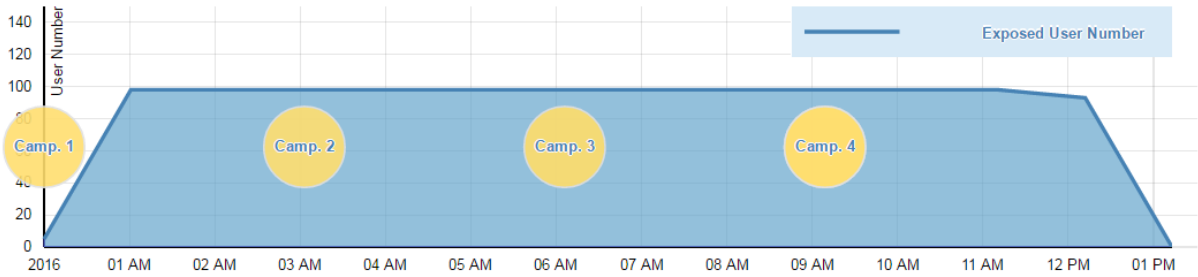


Figure 3.8 Campaign Design with Reduced Campaign Interval

VI. Conclusions and discussion

This paper is targeting to help social media campaign designers to filter out potential users through topic-specified user's behavior study which involves text mining, users' clustering, and optimization.

An interactive visualization tool is developed based on the social network information diffusion simulation and human memory theory. Equipped with the interactive visualization tool, campaign designers can tune their campaign parameters such as rounds of campaign and campaign time intervals before launching an expensive campaign.

In this paper, we classify all followers into "active followers" and "passive followers" based on their interactive retweet times. However, a more precise model can serve to predict users' behavior with an enhanced accuracy. Currently, we assume all active followers of a specific influencer will retweet once they are exposed to the campaign information, which is an over optimistic model for the information diffusion. In the future, we may incorporate a stochastic model to describe the tweet & retweet rate among twitter users. This will allow us to calculate the retweetability of each user and make our simulation more realistic and convincing.

VII. Appendix: Team member task effort

All team members contributed a similar amount of effort.

	Data Processing	Algorithm Design	Visualization Tool	Technical Writing
Jiajie	✔	✔		✔
Yue		✔	✔	✔
Chia-Jung	✔		✔	✔

Yingjie	✔		✔	✔
Xiaochen		✔	✔	✔

VIII. Reference

[1] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, “Everyone’s an influencer: Quantifying influence on twitter,” in Proceedings of the fourth ACM international conference on Web search and data mining. ACM New York, NY, USA, 2011, pp. 65–74.

[2] Ethem F. Can and Hüseyin Oktay and R. Manmatha, “Predicting ReTweet Count Using Visual Cues”, 2013, San Francisco, CA, USA.

[3] E. Katz and P. F. Lazarsfeld. “Personal influence; the part played by people in the flow of mass communications”. Free Press, Glencoe, Ill.” 1955.

[4] Meeyoung Cha, Hamed Haddadi, Fabr´ıcio Benevenuto and Krishna P. Gummadi. “Measuring User Influence in Twitter: The Million Follower Fallacy”, AAAI, 2010.

[5] O. Aarts, P. P. van Maanen, T. Ouboter and J. M. Schraagen, "Online Social Behavior in Twitter: A Literature Review," Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on, Brussels, 2012, pp. 739-746.

[6] Parikh, Ravi, and Matin Movassate. "Sentiment analysis of user-generated twitter updates using various classification techniques." CS224N Final Report(2009): 1-18.

[7] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." the Journal of machine Learning research 3 (2003): 993-1022..

[8] Rendle, Steffen, et al. "Fast context-aware recommendations with factorization machines." Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. ACM, 2011.

[9] Puigbo, Jordi-Ysard, et al. "Influencer Detection Approaches in Social Networks: A Current State-of-the-Art." CCIA. 2014.

[10] B. Suh, L. Hong, P. Pirolli, and E. H. Chi, “Want to be retweeted? large scale analytics on factors impacting retweet in twitter network,” in Proceedings of Social Computing / IEEE International Conference on Privacy, Security, Risk and Trust, 2010.

[11] Willis, Alistair, Ali Fisher, and Ilia Lvov. "Mapping networks of influence: Tracking Twitter conversations through time and space." Participations: Journal of Audience & Reception Studies 12.1 (2015): 494-530.

[12] Cossu, Jean-Valère, Vincent Labatut, and Nicolas Dugué. "A Review of Features for the Discrimination of Twitter Users: Application to the Prediction of Offline Influence." arXiv preprint arXiv:1509.06585 (2015).

[13] G. Ver Steeg, R. Ghosh, and K. Lerman, “What stops social epidemics?” in Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, Association for the Advancement of Artificial Intelligence. AAAI Press, Menlo Park, California, 2011, pp. 377–384.

[14] Seidman, Stephen B. "Network structure and minimum degree." Social networks 5.3 (1983): 269-287.
[15] Hao, Ming, et al. "Visual sentiment analysis on twitter data streams." Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on. IEEE, 2011.