MITBBS Project Final Report

Group Member:

Yawen Lei
Yang Liu
Xiandan Zhou
Beilei Zhang
Introduction

Motivation
MITBBS is one of the most influential online forums in the oversea Chinese community. One characteristic that distinguishes it from other Chinese online forums is that it has no censorship imposed by the Chinese government. That makes it an interesting object to study how Chinese people participate in online political discussions. This is also our first attempt in the broader research agenda on how political systems, in particular, speech regulations, and democracy experiences, influence the quality of online political participation, and how to use technology to promote freedom of speech in China. In this study, we will investigate political participation in three boards: "Military", "Salon", and "ChinaNews" since the three boards are the main forums for political discussion in MITBBS and it seems to us that most of the participants in the three boards have distinct ideologies -- "Military" for patriots, "Salon" for dissents, and "ChinaNews" for independents.

Research questions
We aim to address the following questions in this study.

1) Do participants in the three boards only communicate with participants with similar ideology or also communicate with participants with distinct ideology? The outcome could look like the Democratic vs. Republican graph in Adamic and Glance (2005). But knowing participants’ ideology is a difficult task. There might be two ways to identify ideology. First, we can assume that participants of ChinaNews, Military, and Salon boards are quite ideologically homogenous within the groups while heterogeneous between groups according to our preliminary online observation. Alternatively, we can do content analysis to investigate participants’ ideologies.

2) Do the three boards have similar network structure, in particular, centralization? The overall centralization can tell us if a board is controlled by few participants.

3) Are the posts in the three boards informative? By informative we mean that participants’ posts are more than simple agreement, disagreement, or personal attack. Also, we are wondering if there is any relation between informativeness and the centrality of participants in the MITBBS "Military"+"Salon"+"ChinaNews" combined network and the three board networks separately.

4) Can we detect any community structures in the MITBBS "Military"+"Salon"+"ChinaNews" combined network?
Literature Review

We’ve reviewed two groups of literature. The first group is more social, humanistic oriented, and the second group is more information technology oriented. We are mainly motivated by the first group of literature and seek technical solutions from the second group of literature.

Social, political literature

In the body of literature on deliberate democracy dated back to Habermas’s Structural Transformation of Public Sphere (1989), political theorists emphasize the significance of the quantity (i.e., the openness of public sphere) and the quality of discourse. According to some lines of social and political theories, critical/rational, and cross-ideological conversation could enhance understanding of people with different identities and stances, and is consequently significant in facilitating just societal decisions. Because of the potential progressiveness of conversation, we are motivated to analyze the structures and content of conversation in MITBBS.

Given the importance of conversation on democracy, study of whether conversations occur within or/and between ideological groups, as well as how informativeness conversations, would be meaningful. Prior literature debates over if the Internet facilitates dialogue across ideology boundaries or cause polarization (Sunstein). Although there are some social empirical studies in the US context (Hargittai 2008), studies of political discussions in the Chinese context is lacking. Also, the investigation of cross-ideology linking and discussions in the political blogosphere is not realistic in the Chinese context. Since bloggers encounter tremendous political pressure and risks to run political related blogs, discussions about political affairs usually occur in the discussion forums rather than blogsphere. This motivates us to study the interaction in the Chinese political forums rather than the blogsphere.

Information technology literature

To study online political discourse, researchers engaged in a novel method -- social network analysis. Adamic and Glance (2005) studied link behaviors in the blogosphere during the 2004 presidential election, and found that political blogs rarely link to other blogs with opposing political view. Kelly et al (2005) found that considerable reciprocity existing among online users, especially when they have different opinions. Kou et al (2003) and Goh et al (2006) explored different ways of constructing a network from discussion threads. Those prior researches inspired us to do similar study, by using a combination of methods in the previous studies, on the Chinese online political discussions.

We also reviewed literatures about online discussion and network analysis in general. Smith (2002) talked about the Netscan tool to study large-scale forums. Fiore et al (2002) discussed some general properties of online discussions. Several papers by Newman (2004, 2006) talked about network community finding algorithms. We borrowed some ideas from those researches too, and applied them in this project.
Data Collection

Data resources
MITBBS provides both Web interface access and a more efficient Telnet terminal access. We wrote a Python program to scraped data of 3 major political boards from the Telnet interface once per day from 2008-9-1 to 2008-10-9. The political boards of MITBBS are:

1. Military: a board for patriots who are supporters of the Chinese government;
2. Salon: a board for liberals who usually hold the anti-government position;

After 39 days of data collection, we had a total number of 132468 posts in our database, about 43% of them from ChinaNews, 48% from Military, and 9% from Salon.

Data selection
Since current network analysis tools such as GUESS or Pajek don’t really support a large network dataset such as ours, which would have 4291 author IDs as nodes in the network, we reluctantly restricted our attention to only a small subset of data. After making a post distribution graph for those data, we finally chose all posts in 9/13/2008 as our dataset, because there was a sudden spike of discussions in that day. There were about 5868 posts and 527 threads. We build our network model on that subset of data, and the rest of the discussion is also based on this data.

Figure 1. Posters distribution from 09/01/2008-10/09/08

Figure 2. Log-log plot of total number of posts per author
Network Construction

Construct network from discussion threads

The first question we encountered was how to represent the network. We treated each author as a node in the network. If there was some form of interactions between the authors, we would draw a link between the nodes. Our problem is how to define an interaction. In previous works (Kelly et al, 2005), there would be a directed link from B to A if B responded to A’s previous post. However, our data has a limitation that we cannot track down the exact response relationship. All posts in the same thread are listed in a chronological order, usually without explicitly pointing out to which posts they intended to respond. After looking into literature, we found that we have three choices—line, star, and complete—to construct the network approximately.

The “line approximation” is based on the assumption that everyone in the thread just responses to their immediate preceding post, which is not likely based on our observation. The “star approximation” gives excessive emphasis on the thread starters. In our study, we adopted the “complete approximation” to construct the network, where the edges have different weights based on their relative importance in the thread. Intuitively, suppose author A made 1 post in the thread whereas both B and C made 10 posts in the same thread, then the link between B and C would have the highest weight. In our implementation, we gave an equal amount of scores to each thread, and distributed the scores evenly to each post in the thread. Since each post has a single author ID, we then aggregated the scores of the posts by authorship, and used the sum of the two authors’ scores as the weight of the interaction between the two authors in the same thread. Finally, we aggregated the weights of all pair-wise interactions between authors to be the final weight of edge between two nodes in the network. We then generated 3 network files, GDF, GraphML, and Pajek for further analysis using different network tools.

Note that the network constructed this way is an undirected weighted network. That limits our ability to study reciprocity, or strongly connected components, which can only be performed on directed network. We will address this problem in future works.
As we have mentioned in the research questions, we need to identify participants' "informativeness" and ideology to conduct further analysis. We tried to rate each post in terms of "informativeness" and "political position".

First, we measure how much relevant information contained in the posts. We coded the informativeness with the following 5-point scale. After rating all of the posts, we calculated the average "informativeness score" for each node in the network.

### Table 1: Informative scale

<table>
<thead>
<tr>
<th>Scale</th>
<th>Informative</th>
<th>Explain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Irrelevant</td>
<td>Totally irrelevant with meaningful political discussion, such as personal attack.</td>
</tr>
<tr>
<td>2</td>
<td>Almost no information</td>
<td>Such as simply express agreement or disagreement without any explanation.</td>
</tr>
<tr>
<td>3</td>
<td>Limited information</td>
<td>Give limited reasons of their disagreement or agreement, or add limited extra information albeit not explicitly expressing their opinions.</td>
</tr>
<tr>
<td>4</td>
<td>Moderate information</td>
<td>Give moderate information.</td>
</tr>
<tr>
<td>5</td>
<td>Abundant information</td>
<td>Posters with very complete and articulated opinions.</td>
</tr>
</tbody>
</table>
To ensure the objectiveness of rating, we planned to have two team members rating the first 500 posts and to see if we can have reasonable inter-rater reliability. If the reliability is okay, we will train the computer with some machine learning algorithm to rate the rest of the posts. If not, we will have the third rater.

In the rating process, we built a Drupal web interface to facilitate two group members rating the posts, as shown in the figure below. However, we found that the Cohen's inter-rater coefficient for the first 500 posts by two raters was only 46% for "informativeness" and 95% for "political position". It means the two raters didn't have much agreement on the rating of "informativeness". We thought it was because it takes a lot of time to cultivate a "stable" feeling of rating. But since we didn't have enough time to have the third person rating, we solely relied on one member's rating of "informativeness" to conduct further analysis. (Note: we will address Lada's comment on alternative rating scale in the Appendix.)

![Drupal Interface](image)

Figure 4: Drupal Interface

With regard to "political position," we wanted to measure whether a node supports current Chinese government or against it by appraising the posts of each node. We have 4 categories:

Table 2: Rating of political position

<table>
<thead>
<tr>
<th></th>
<th>Political Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>For Government</td>
</tr>
<tr>
<td>2</td>
<td>Neutral</td>
</tr>
<tr>
<td>3</td>
<td>Against government</td>
</tr>
<tr>
<td>4</td>
<td>Information is insufficient</td>
</tr>
</tbody>
</table>
Although we had a very high inter-rater coefficient (95%) for the rating of “political position,” after reviewing all the posts, we found in 99% of the cases, we have inadequate information to tell whether a person is against government or not. Many posts are not directly related to attitude toward government. In many situations, we can tell that online participants are dissatisfied with the Chinese government, but it is hard to tell if they are against the government. Although the results are frustrating, we notice that there are relatively clear divisions in attitudes towards the Western democratic and legal regimes. Hence, it could have been much better if we see whether a node is "pro democracy" or "against democracy". Since the measurement we used in rating was not good, we had to discard the measurement on "political position". In the future, we will need to revise our operationalization of the political position measurement. In further analysis, we can only assume that participants of ChinaNews, Military, and Salon boards are quite ideologically homogeneous within the groups while heterogeneous between groups according to our preliminary online observation.
Descriptive Statistics

The "Military"+"Salon"+"ChinaNews" network we constructed has 925 vertices and 30733 edges. Its density is around 0.07.

Table 3: Basic Descriptive Statistics of the "Military"+"Salon"+"ChinaNews" network

<table>
<thead>
<tr>
<th></th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post #</td>
<td>925</td>
<td>6.343784</td>
<td>15.14864</td>
<td>1</td>
<td>302</td>
</tr>
<tr>
<td>Inform</td>
<td>925</td>
<td>2.868828</td>
<td>0.795554</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Closeness</td>
<td>925</td>
<td>0.4562754</td>
<td>0.0765315</td>
<td>0</td>
<td>0.6614222</td>
</tr>
<tr>
<td>Betweenness</td>
<td>925</td>
<td>0.0012133</td>
<td>0.0044475</td>
<td>0</td>
<td>0.0789698</td>
</tr>
<tr>
<td>Degree (normalized)</td>
<td>925</td>
<td>0.0719153</td>
<td>0.0718934</td>
<td>0</td>
<td>0.5054113</td>
</tr>
</tbody>
</table>

In the "Military"+"Salon"+"ChinaNews" network, we did not find strong and consistent evidence on the correlation between informativeness and three centrality measurements after calculating both Pearson Correlation and Spearman’s Rank Correlation. That means that participants with high degrees, betweenness, or closeness are not more informative or less informative than those with lower centrality.

Table 4: Pearson Correlation between Variables

<table>
<thead>
<tr>
<th></th>
<th>Post #</th>
<th>Inform</th>
<th>Closeness</th>
<th>Betweenness</th>
<th>Normalized degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post #</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inform</td>
<td>-0.0436</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Closeness</td>
<td>0.343</td>
<td>-0.1082</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Betweenness</td>
<td>0.8108</td>
<td>-0.0536</td>
<td>0.3506</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Normalized degree</td>
<td>0.5915</td>
<td>-0.0606</td>
<td>0.6711</td>
<td>0.6618</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 5: Spearman’s Rank Correlation

<table>
<thead>
<tr>
<th></th>
<th>Post #</th>
<th>Inform</th>
<th>Closeness</th>
<th>Betweenness</th>
<th>Normalized degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post #</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inform</td>
<td>0.0094</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Closeness</td>
<td>0.573</td>
<td>-0.0231</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Betweenness</td>
<td>0.8704</td>
<td>-0.0717</td>
<td>0.6681</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Normalized degree</td>
<td>0.5327</td>
<td>-0.0234</td>
<td>0.9628</td>
<td>0.6228</td>
<td>1</td>
</tr>
</tbody>
</table>
We then looked at if participants with different main membership are different in terms of the number of posts and their informativeness. Each node is assigned a main membership according to the maximum number of posts he or she has in the three boards. For instance, if a person has one, three, and ten posts in ChinaNews, Military, and Salon, respectively, his main membership is Salon. From the table, we can see that (do we need to mark each table with numbers?) Participants with the ChinaNews main membership tend to have more posts, while participants with the Military main membership tend to have fewer posts.

Table 6: Average Number of Post Per Participant

<table>
<thead>
<tr>
<th>Board</th>
<th>Number of Post (Mean)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChinaNews</td>
<td>7.3944444</td>
</tr>
<tr>
<td>Military</td>
<td>4.3812709</td>
</tr>
<tr>
<td>Salon</td>
<td>6.5697674</td>
</tr>
</tbody>
</table>

We compare the informativeness of people with different memberships and found that participants with the ChinaNews main membership are significantly less informative.

Table 7: Average Informativeness Per Participant

<table>
<thead>
<tr>
<th>Board</th>
<th>Informativeness (Mean)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChinaNews</td>
<td>2.9137063</td>
</tr>
<tr>
<td>Military</td>
<td>2.7707151</td>
</tr>
<tr>
<td>Salon</td>
<td>2.9281445</td>
</tr>
</tbody>
</table>
Findings

Comparison of centralization of the three boards

The overall degree centralization of the "Military"+"Salon"+"ChinaNews" network is 0.43444. But the three boards are quite differently centralized. Salon has the highest degree centralization; Military has the lowest degree centralization. This suggests that Salon is dominated by relatively few participants. In contrast, participation in Military is more evenly distributed. The following three graphs demonstrate this difference.

Table 8: Centralization of three Boards

<table>
<thead>
<tr>
<th>Board</th>
<th>Degree Centralization</th>
<th>Betweenness Centralization</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChinaNews</td>
<td>0.56315</td>
<td>0.07441</td>
</tr>
<tr>
<td>Military</td>
<td>0.47313</td>
<td>0.09342</td>
</tr>
<tr>
<td>Salon</td>
<td>0.64305</td>
<td>0.19814</td>
</tr>
</tbody>
</table>

Figure 5: Sized by degree network for ChinaNews-> Military-> Salon

The Relations between Informativeness and Centrality

We are also interested in the relation between informativeness and centrality. Being informative is usually appreciated in the conversation since informativeness contributes to mutual understanding and represents some kind of ability to influence other people. If centrality can mean positive reputation, then there could be positive correlation between informativeness and centrality. On the other hand, given the limited time each participant has, being central could mean that participants are involve in too much discussion, so their discussion is not so informative. If it is true, then there could be negative correlation between informative and centrality.
First we examine the correlation between informativeness and centrality in the "Military"+"Salon"+"ChinaNews" network. It turns out that the two variables are quite independent in the Military"+"Salon"+"ChinaNews" network. But considering that there could be different "culture" in the three boards as the three boards are organizationally differentiated, we also looked at the relation between informativeness and centrality in the three boards respectively. We found that there is small negative correlation between the two variables in the ChinaNews board, and small positive correlation between the two variables in the Military board. However, there is still no significant relation in the Salon board.

<table>
<thead>
<tr>
<th>Table 9: Correlations between Informativeness and Centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Pearson Correlation</td>
</tr>
<tr>
<td>Spearman's rho</td>
</tr>
<tr>
<td>p-value of Spearman test</td>
</tr>
</tbody>
</table>

How can we interpret these results? We think the composition of the participants might provide a clue. Participants in the ChinaNews board are with relatively diverse main membership. The number of node is also larger. It is not a very cohesive community with a strong sense of "we-ness." It probably doesn't matter too much if you say something informative or not. Also, the topics discussed in ChinaNews tend to be more general. When the sense of community does not prevail, given the time constraints, the more participants involved in discussion, the less informative their posts are.

By contrast, participants in the Military board are more homogeneous in terms of their main membership. In a cohesive community, participants might have more incentive to be informative to gain respect from other participants. Salon board is more heterogeneous than military but not so much as ChinaNews in terms of the main membership of participants. Some participants in the peripheral area of Salon network are of ChinaNews and Military main membership. In the process of content analysis, we found that the intention of those participants is to challenge the opinions of participants with the Salon main membership. Because they intend to win the battles, they are usually articulated even though they don't participate so often in Salon as those with Salon main membership do. Thus, the informativeness of participants in the core and in the periphery doesn't vary a lot.

These results tell us that the relations between informativeness and centrality might be influenced by social conditions.

Community Structure
Next, we studied the community structure of the network. Due to size of the network (over 900 vertices), we tried to detect the community structure via two methods, including hierarchical clustering and greedy optimization of modularity.

Hierarchical clustering in Pajek
First method is to use the hierarchical clustering method in Pajek which is based on the traditional edge-independent path counts algorithm. The dendrogram graph is attached in the appendix.

The dendrogram shows the hierarchical structure of the network. As we want to divide the network into three communities, we cut the slice at the level which is marked by the red line in the dendrogram. Because of the large number of the vertices, the dendrogram is not that straightforward to detect the community structure. More importantly, due to the drawback of the algorithm itself, the distribution of vertices in three communities is highly skewed that over two thirds of the vertices form a dominant community. This has little correlation with the actual split of the members of three boards.

Greedy optimization of modularity
Second method is to detect the community structure via directly optimizing a modularity score (ref 1, 2) with R. We use the fast greedy community detecting function in R to analyze the community structure via greedy optimization of modularity. We get the merge matrix and a vector containing the modularity after each search. Because it takes 911 steps to merge all nodes in one community and we expect to divide the network into three communities, we use community.to.membership(g, fastg$merges, steps=909) to create a membership vector by performing a given number of merges in the merge matrix. The community size (909 steps) is described as the following vector:

$csize
[1] 299 300 312 2 1 1 1 1 1 1 1 1 1 1 1

The vector shows that the community 0 contains 299 members, community 1 contains 300 and community 2 contains 312 members. The fourth and the rest have only one or two nodes. The following is the membership vector which shows each node's community label.

[1] 0 1 1 0 0 2 2 2 1 2 0 1 0 2 2 1 2 1 2 0 2 4 1 3 2
[26] 0 1 2 1 2 2 0 2 1 0 0 1 2 1 1 0 1 2 0 2 2 2 1 2
[51] 2 1 2 2 0 0 0 2 0 0 0 1 1 1 1 2 0 2 1 2 2 1 0 2 1
[76] 5 2 0 1 2 1 2 2 1 0 1 2 2 1 2 1 1 0 2 1 1 1 1 1 2
[101] 0 0 2 1 1 0 2 1 1 1 0 2 1 2 1 0 2 1 0 0 1 2 0 0
[126] 1 2 0 0 2 6 0 0 1 1 2 2 0 2 0 0 2 0 2 1 2 2 0 2 0 1
[151] 0 1 0 0 1 2 1 1 1 2 0 2 2 2 2 2 2 2 1 1 2 0 1
[176] 0 2 2 0 0 2 2 0 1 2 2 1 1 1 1 1 0 0 2 2 1 1 0 0
[201] 2 1 2 1 0 2 0 2 1 2 1 1 2 1 1 2 1 1 1 1 1 2 0 1 0
[226] 1 1 0 0 2 1 0 2 0 2 0 2 0 2 2 1 0 0 0 2 0 0
[251] 3 1 1 2 0 0 2 2 1 2 0 2 1 0 1 1 1 1 0 0 1 1 2 0 1
[276] 2 1 2 2 0 1 2 2 1 1 1 1 0 0 0 0 2 2 0 2 1 2 0 2 0
[301] 2 2 1 1 2 1 0 1 1 0 2 2 0 2 0 1 2 1 0 2 1 0 2 2
[326] 1 2 1 0 1 2 1 2 2 1 2 2 1 2 1 0 1 0 1 1 1 0 2
[351] 0 2 2 1 1 2 1 1 1 2 0 2 1 1 0 2 1 0 2 2 0 2 0 0
[376] 0 2 2 1 0 1 0 2 2 2 1 2 1 1 1 2 2 0 2 0 2 0 0
[401] 0 1 2 0 0 2 2 1 0 0 2 0 2 2 2 1 1 0 1 2 1 0 0
[426] 1 2 1 2 0 0 2 0 1 2 0 0 0 2 1 1 1 2 1 2 1 2 2
[451] 0 7 8 2 1 1 0 0 0 1 2 0 0 0 0 9 0 0 1 2 0 0 1 1
[476] 2 1 1 0 2 0 1 2 2 2 0 1 2 2 1 0 2 0 1 2 1 0 2 0 2
[501] 1 1 0 2 0 0 2 1 1 1 1 2 0 0 2 1 2 0 1 2 1 2 0 0
[526] 1 1 0 1 2 1 1 0 1 2 0 1 0 1 2 2 1 1 2 0 1 1 2
[551] 0 1 0 2 0 1 0 2 2 0 0 0 0 0 1 0 1 0 1 0 1 0
[576] 1 0 2 1 2 2 1 2 0 2 0 2 2 0 2 1 2 0 2 1 0 1 0 0 1
[601] 2 1 0 0 1 2 0 2 1 1 2 2 2 1 2 2 1 2 2 0 0 0 1 0 2
[626] 0 1 0 1 1 1 0 2 0 1 1 1 2 2 1 2 1 1 0 1 2 2 0 1
[651] 0 2 1 0 2 0 1 0 2 1 2 2 0 2 2 2 2 0 2 1 2 0 1 2 0 0 1 1
[676] 0 0 0 0 0 2 1 1 2 2 0 0 0 2 0 2 2 2 1 0 1 2 1 1
[701] 0 1 0 1 1 2 2 2 0 2 2 1 2 1 0 1 2 1 1 1 0 2 1 0
[726] 2 2 2 0 2 1 2 0 1 2 2 0 1 2 2 2 2 0 1 2 2 0 1 1 1
[751] 2 2 2 0 2 1 2 1 0 1 2 1 0 2 1 0 1 2 2 2 0 1 1 1
[776] 0 2 0 1 1 1 0 2 0 0 2 0 2 0 1 0 0 1 2 0 1 0 1 3 2 1 0
[801] 1 0 2 0 1 0 1 2 1 1 0 2 2 1 2 1 2 1 1 0 1 1 2 1 2
[826] 2 0 2 0 2 1 0 2 0 2 1 2 0 2 1 2 0 1 0 2 2 2 0 2
[851] 0 0 1 4 1 1 1 0 1 1 0 1 1 2 2 1 2 1 0 1 2 0 1 1 0
[876] 0 2 0 2 1 5 1 1 1 0 2 0 2 0 1 0 0 2 0 1 2 1 2 0 1
Finally we attach each membership label to the node information as an attribute called "community_id" in the data file.

**Visualization of community structure**

We visualize the network with GEM layout. Then we colorize the vertices according to vertices' community membership and main membership respectively.

**Figure 6:** Colorization according to community: Red-1; Yellow-2; Cyan-0

**Figure 7:** Colorization according to main membership: Red-Military; Yellow-ChinaNews; Cyan-Salon
By comparing two visualizations, we can see that the distribution of the red nodes have a great correlation in two graphs. This shows that members of Military board prefer communicating only within their own group. However, the distribution of yellow and cyan nodes are not that identical which can be interpreted as the communication between members from ChinaNews and Salon boards. To understand the composition of each community more accurately, we look into each community to see how they are composed by three board members.

Table 10: Composition of three communities

<table>
<thead>
<tr>
<th>Board community</th>
<th>ChinaNews</th>
<th>Military</th>
<th>Salon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Community 0</td>
<td>197</td>
<td>19</td>
<td>82</td>
</tr>
<tr>
<td>Community 1</td>
<td>28</td>
<td>272</td>
<td>0</td>
</tr>
<tr>
<td>Community 2</td>
<td>308</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

From the pie charts, we can see that community 1 is mainly composed of participants with Military Board main membership. Very few members whose main membership is Military are in the other two communities. That means these people interact mostly with each other instead of interacting with other two boards members. This result is consistent with our interpretation in 6.2 that the Military board is a relatively cohesive community. Participants with ChinaNews main membership mainly separate into two parts. About 37% of them form community 2 themselves. Another 57% of them merge into community 0 with almost all members from Salon Board. Instead of forming an independent community themselves, the participants with Salon main membership form a community with some ChinaNews board members. They post sometime in ChinaNews besides post in Salon. Although we could not examine participants’ ideology directly, the result of community structure detection suggests that participants with the ChinaNews main membership tend to communicate with each other. According to our preliminary observation, ChinaNews participants seem to have a pro-government ideology. Thus, it is fair to claim that participants who support the Chinese government tend to interact with each other and form a cohesive community on the MITBBS.

Figure 8: Pie charts of composition of three communities
Conclusion

In this study, we attempted to explore several questions. First, we’d like to see if participants in the three boards only communicate with participants with similar ideology or also communicate with participants with distinct ideology, but it turned out that we could not identity participants’ ideology directly. On the other hand, our result of community structure detection suggests that participants who support the Chinese government tend to interact with each other and form a cohesive community on the MITBBS. Second, we found the Salon board is more centralized, meaning that it is controlled by relatively few participants. Third, we found that there is no correlation between informativeness and degree centrality in the "Military"+"Salon"+"ChinaNews" combined network. However, there is negative correlation between informativeness and degree centrality in the ChinaNews network, there is positive correlation between the two variables in the Military network, and there is no correlation between the two variables in the Salon board network. This different pattern of correlation between informativeness and centrality could be due to the different compositions of participants in the three boards. Finally, we found three communities in the "Military"+"Salon"+"ChinaNews" combined network. Participants with the Military main membership have a strong preference to communicate with each other. However, Participants with Salon and ChinaNews main membership tend to interact with both participants with Salon main membership and ChinaNews main membership.
Course Reflection

Working on the MITBBS project helped us to a better understanding of the concepts learned from this course. In this last section, we’d like to explain the setbacks and reflect on the lessons we’ve learned from the process. The first lesson was that why our research question about polarization didn’t work. This was due to the fact that we didn’t have an accurate intuition about what the posters talked about before doing the study. As a result, we picked a wrong measurement which didn’t work well in the rating process. The next lesson was that machine classifier and sentiment analysis was not mature enough to be easily applied in serious research. A good research of this kind probably requires industrious content analysis and qualitative method. In short, if we have had done a pre-study before the semester and carefully design the research before diving into the work, the outcome could have been improved a lot.

But, in fact, this project, with its highly exploratory nature, can be considered as a pre-study itself. We hope to use what we’ve learned here to start a broader research agenda about Chinese political online deliberation. Particularly, we hope to continue the research in these possible directions. The first is to learn more about the properties of Chinese online political discussions, such as reciprocity, the degree of polarization, the spectrum of distribution of political ideology, and so on. That could be done by constructing a directed network using the method discussed in (Goh et al, 2006), which requires a more sophisticated program that generates the "reply to" metadata. It also requires better measurement scale, and probably a functional auto-classifier or content analysis by human readers. The second direction is to extend the study to online political forums in mainland China, Taiwan and oversea, and see if there are differences and how they differ. This requires a set of metrics that measure the achievement of each forums in various dimensions, such as reciprocity, diversity, the degree of polarization, and so on. The third direction is about how oversee Chinese people, by participating in unrestricted online discourses, change their political opinions and beliefs. The last direction is to base on what we have learned from and study and make design changes to online political discussion that could achieve better outcomes.
References

14. igraph document
Appendix:

A.1. Addressing reviewer’s feedback

**Mid-project feedback** -- I just want to point out that the higher closeness, betweenness, degree of the ChinaNews forum could be due entirely to the higher number of posts/author in that forum. What would be more informative is the “centralization” rather than the average centrality. That is, to what extent do a few individuals dominate the discussion in each of the forums?

**Response:** We addressed this issue in *Comparison of centralization of the three boards* in the Findings section.

**Mid-project feedback** -- I think you could do a bit more to compare the forums to one another, perhaps looking at the size of the largest strongly connected component, the degree of reciprocity, etc.

**Response:** To address this feedback, we compared the centralization of the 3 forums, and made visualization of each of them. However, one limitation of our network is that it is constructed as a “undirected network” (see Network Construction section for details). Thus we are unable to calculate the strongly connected component or the degree of reciprocity. But, of course, those are very good comments we should address in future works.

**Mid-project feedback** -- BTW, when you say that the community alignment proves your hypothesis, which hypothesis is that?

**Response:** In fact, we didn’t have particular hypothesis and have deleted the wordings about hypothesis.

**Mid-project feedback** -- For the difficulty in getting reliable informativeness ratings, I would suggest using a 3 point scale rather than a 5 point one: 1. useless, 2. express opinion but no justification, 3. includes justification

**Response:** This is a very good suggestion. We should have used this measurement from the beginning. However, human coding is very expensive (it took us a whole week rating those in the first round), and we don’t have enough resources to rate it again. Instead, due to the limited scope of this project, we think it’s fine just to trust the results from a single rater. We’ll adopt more appropriate measurement in future research based on the experience we learned here.

**Mid-project feedback** -- As for the polarization ratings, I trust you know what you’re doing :)

**Response:** We only got to learn the measurement for polarization ratings were not appropriate during the work. And again, we don’t have enough resources to go back and rate all the posters again. We’ll adopt more appropriate measurement in future research.

**Final presentation feedback** -- To what extent is the community structure an artifact of creating complete cliques for each thread?

**Response:** This is a valid comment. Our best guess is that the community structure replies a lot on the “complete” network construction for each thread, because this makes the all participants in the thread close to
each other. We would reasonably think that the "line" or the "star" network construction for each thread would yield different network structure. It would be interesting though to generate 3 networks, "line", "star" and "complete" and verify whether the community structure is different or not. But due to our limited resource, we are not able to do it in this work. It will be addressed in future works.

A.2. Programs and scripts

The Python/PHP program that scrapped the data, generated the network file, recorded human ratings, and computed communities consist of 800 lines of code. For the source code, please contact Daniel Zhou at mrzhou@umich.edu