Community Detection on a Social Network[*]

Jieyu Gao
Department of Economics
University of California-Irvine
jieyug1@uci.edu

Hao-Che Hsu
Department of Economics
University of California-Irvine
haoche.hsu@uci.edu

Hanqiao Zhang
Department of Economics
University of California-Irvine
hanqiaoz@uci.edu

Abstract

Working with network data from Hornet social platform, this project compares different community detection approaches including Mixed Membership Stochastic Blockmodels and K-means clustering on node edges and individual demographic attributes. For the stochastic model, the likelihoods of different sampling methods are examined. Finally, the network topology is visualized with Gephi.

1 Background

When observing connections between entities, we like to be able to recover the underlined structure. The structure is assumed that people belong to different communities and similarities are characterized by certain behaviors that allocate them into the same cluster.

Community detection has been an important topic in analyzing social networks, especially in the dating market. Using the data from the Hornet network, we would like to detect the overlapping communities to explore these specific social networks using a mixed-membership stochastic blockmodel. The communities detected in this case may help reveal some connection preferences of different types of users.

Using the Mixed Membership Stochastic method proposed in [1] on the top 5000 nodes, we estimated that there exist 76 communities. We also found that the demographic attributes, including the locations and languages, play important role in forming different communities.

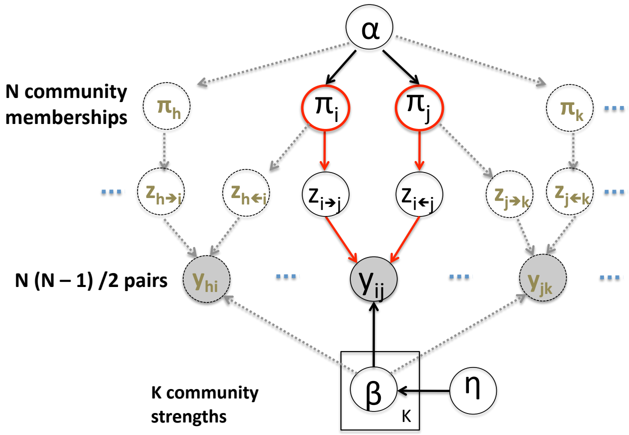
2 Mixed Membership Stochastic Blockmodels

Unlike traditional data, community information contains interrelated observations. However, standard clustering methods such as mixture models which rely on conditional independent observation given their cluster assignments can be misspecified.

[*] ID: Jieyu Gao 407-221-18, Hao-Che Hsu 480-474-20, Hanqiao Zhang 576-064-12.

CS 274B: Graphical Models Course Project (2021).

Figure 1: a-MMSB [2]



[3] proposed the latent stochastic blockmodel which allocates each entity to a cluster. The relationships between objects are determined by the corresponding pair of clusters. Nevertheless, the immediate backslash from this assumption of limiting each object can only belong to a single cluster or to say, play a single latent role, to some extent, unrealistic since in a large social network people might belong to overlapping communities.

Rather than a single cluster, based on the development of the mixed membership model for relational data which associates each unit of observation with multiple clusters by a membership probability vector or a membership matrix, the MMSB [4] assumes the membership distributions of the entities are independently drawn from a Dirichlet distribution. [1] and [2] extend the model to large observed networks by using the assortative mixed-membership stochastic blockmodel (a-MMSB), as shown in **Figure 1**.

The model assumes that if individuals i and j are connected, they have at least one common community assignment. From there, there are K communities, and each node i is assigned a vector of community membership θ_i. Thus, the probability that there is a link between nodes i and j is indicated as follows:

$$p(y_{ij} = 1 \big| \theta_i, \theta_j) = \sum_{k=1}^{K} \theta_{ik}\theta_{jk}\beta_k \tag{1}$$

where β_k represents how densely the community k is. Due to the difficulties that arise when calculating the posterior distribution directly, they proposed a way to estimate the model using variational inference. The mean-field variational family is defined as follows:

$$q(\beta, \theta, z) = \prod_{k=1}^{K} q(\beta_k|\lambda_k) \prod_{n=1}^{N} q(\theta_n|\gamma_n) \prod_{i<j} q(z_{i\to j}|\phi_{i\to j})q(z_{i\leftarrow j}|\phi_{i\leftarrow j}) \tag{2}$$

where $q(z_{i\to j} = k) = \phi_{i\to j,k}$, $q(\theta_n|\gamma_n) = \text{Dirichlet}(\theta_n|\gamma_n)$, and $q(\beta_k) = \text{Beta}(\beta_k|\lambda_k)$. The objective is to minimize the KL divergence between the distribution q and the true posterior distribution. Due to the feasibility, this objective function can be solved by optimizing the lower bound function as follows:

$$\begin{aligned} L = &\sum_k \mathop{\mathbb{E}}_q[\log p(\beta_k|\eta)] - \sum_k \mathop{\mathbb{E}}_q[\log q(\beta_k|\lambda_k)] + \sum_n \mathop{\mathbb{E}}_q[\log p(\theta_n|\alpha)] - \sum_n \mathop{\mathbb{E}}_q[\log q(\theta_n|\gamma_n)] \\ &+ \sum_{a,b} \mathop{\mathbb{E}}_q[\log p(z_{a\to b}|\theta_a)] + \mathop{\mathbb{E}}_q[\log p(z_{a\leftarrow b}|\theta_b)] \\ &- \sum_{a,b} \mathop{\mathbb{E}}_q[\log q(z_{a\to b}|\phi_{a\to b})] - \mathop{\mathbb{E}}_q[\log q(z_{a\leftarrow b}|\phi_{a\leftarrow b})] \\ &+ \sum_{a,b} \mathop{\mathbb{E}}_q[\log p(y_{ab}|z_{a\to b}, z_{a\leftarrow b}, \beta)]. \end{aligned}$$

The stochastic variational algorithm is described as follows:

At iteration t:

1. Subsample a set of pairs of nodes S.
2. For each pair $(i, j) \in S$, use the current community structure to compute the interaction parameters $\widehat{\phi}_{i\rightarrow j}$ and $\widehat{\phi}_{i\leftarrow j}$.
3. Adjust the community memberships γ.
4. Repeat the process.

3 The Data

Table 1: Hornet Data Descriptions of User Attribute

Variables	**Explanations**
`birth_date`	Time in days when the account is registered
`birth_info_platfor`	Indicator of the platform that is used (IOS or Android)
`birth_info_country`	Country where the user is born
`has_photo`	Indicator of whether the user has uploaded at least one photo
`country_code`	Country where the user register the account
`language`	Language that is spoken by the user
`visible`	Indicator of whether user profile is set to be visible
`lat`	Latitude
`lng`	Longitude
`has_email`	Indicator of whether the user has linked a email address
`feed_post_count_v4,5,6`	Number of feed post made by the user in app version 4,5,6
`feed_engage_count_v4,5,6`	Number of feed post liked or commented by user in app version 4,5,6
`feed_read_count_v4,5,6`	Number of stories read from the feed seen user in app version 4,5,6
`chat_sent_count_v4,5,6`	Number of chat messages sent by user in app version 4,5,6
`guy_follow_count_v4,5,6`	Number of profiles followed by user in app version 4,5,6

3.1 The Hornet Social Network

Founded in 2011, Hornet is the world's premier gay social network. With over 25 million diverse users globally, its mission is to empower gay men to come out and join in the fun and fabulous of the gay community. Hornet has become the number one gay app in markets such as France, Russia, Brazil, Turkey, and Taiwan, and is rapidly expanding its sizable user base in the United States.

The vision is to create a safe space to come out and join the gay community by encouraging people to express their true self where their idiosyncrasies are embraced, and where being open about the things they love is rewarded with rich chat, engaging conversation, and ultimately, meaningful connections.

Before 2017, Hornet was primarily seen as a dating app by users. Since then, the strategy to nudge people into a social network and show them the value of a community is executed more prominently. Since then they launched three Major versions (V4, V5, and V6), each with a vital distinction going from dating to a social network: introducing Hornet Stories – the largest LGBT newsroom back then, changing from Grid to Feed, New profiles, Authentication Badge and introducing a bifurcated inbox.

3.2 Network Data Overview

We follow [1] and [2] to recover the structure of the Hornet network[2]. Hornet is a gay social network that has more than thirty million active users.

The social network dataset contains two tables, the `userSummary` table, and the `connections` table. The `userSummary` table records registration information for 3,246,581 people that have used the social application from July 1st, 2018, to May 10th, 2020. This includes a unique user ID and 25 attributes for each user. The meaning of each user attribute is explained in **Table 1**.

Table 2: Descriptions of User Attribute in the `connections` table

id	**profile_id**	**favourite_id**	**created_at**
58291406	17385489	3038456	5/27/2015 13:26
58016171	17392195	2644369	5/24/2015 14:35
58307179	17393443	16586970	5/27/2015 17:07
57944859	17395904	17432831	5/23/2015 21:08
58516050	17397241	13912544	5/30/2015 2:05

The `connections` table is a snapshot for directed edges that capture the connections among all users, starting from 2011 to June 1st, 2020, the time when this table is generated. A screenshot of the header is shown **Table 2**. Variable id is the unique ID for each edge, profile_id is the user ID for the follower, favourite_id is the user ID for the followee, and created_at records the time for this following event.

3.3 The User Nodes

To understand the degree distribution of vertices, we could start by computing its quantiles as in **Table 3**. It is easy to see that the average and median user degrees are 50.11 and 9. Most users have few connections, and 35% users have less than 5 connections. Connections of the top 1 user are 2.37 times that for the user at the second place.

Table 3: Quantiles of User Degree

0%	**25%**	**50%**	**75%**	**99%**	**99.9%**	**100%**
0	3	9	31	710	2935	164959

Given the distribution of the user degree, a natural guess is whether there are differences among engagement of users from different countries. We grouped the users by their country of registration, calculation the proportions of the user from each country, then rank them within 5 different degree quantiles.

From **Table 4**, it is clear to see that 9 out of the top 10 countries, and 15 out of the top 20 countries stay the same across all quantiles. The top 9 countries are highlighted with bold font. Tier 1 countries are Brazil, Turkey, Indonesia, Thailand. Countries in tier 2 include Russia, the United States, France, Mexico, and Taiwan. Ranks within top countries only make a slight change in different quantiles.

Another thing that could be seen from the degree distribution is that there are in general two types of users in this social network: a few celebrity type users with exceedingly many connections, and regular type users with relatively few connections. Therefore, it is appropriate to divide the users into two groups using the 90% degree quantile threshold (91 connections) and inspect the users in each group respectively.

[2] Hornet: https://hornet.com.

Table 4: Proportions of country of registration among different degree quantiles

0-25%		25-50%		50-75%		75-99%		99-100%	
Country	**Prop**	**Country**	**Prop**	**Country**	**Prop**	**Country**	**Prop**	**Country**	**Prop**
BR	16.89%	**ID**	17.60%	**ID**	17.71%	**TH**	18.91%	**TW**	26.94%
TR	13.98%	**BR**	16.69%	**BR**	16.28%	**BR**	14.32%	**TH**	23.99%
ID	12.23%	**TR**	10.94%	**TH**	12.96%	**ID**	13.52%	**TR**	10.24%
TH	9.69%	**TH**	10.87%	**TR**	10.01%	**TR**	10.65%	**BR**	9.81%
RU	5.73%	**RU**	5.82%	**RU**	5.87%	**TW**	9.07%	**ID**	8.74%
US	4.81%	**FR**	4.20%	**FR**	4.66%	**RU**	5.51%	**RU**	5.05%
FR	4.04%	**US**	3.99%	**TW**	3.39%	**FR**	4.45%	**MX**	2.77%
MX	3.26%	**MX**	3.13%	**US**	3.34%	**MX**	3.54%	**FR**	1.74%
TW	2.27%	**TW**	2.43%	**MX**	3.13%	**US**	2.31%	**UA**	1.19%
GB	1.56%	**MY**	1.59%	**MY**	1.59%	**UA**	1.39%	**US**	1.08%

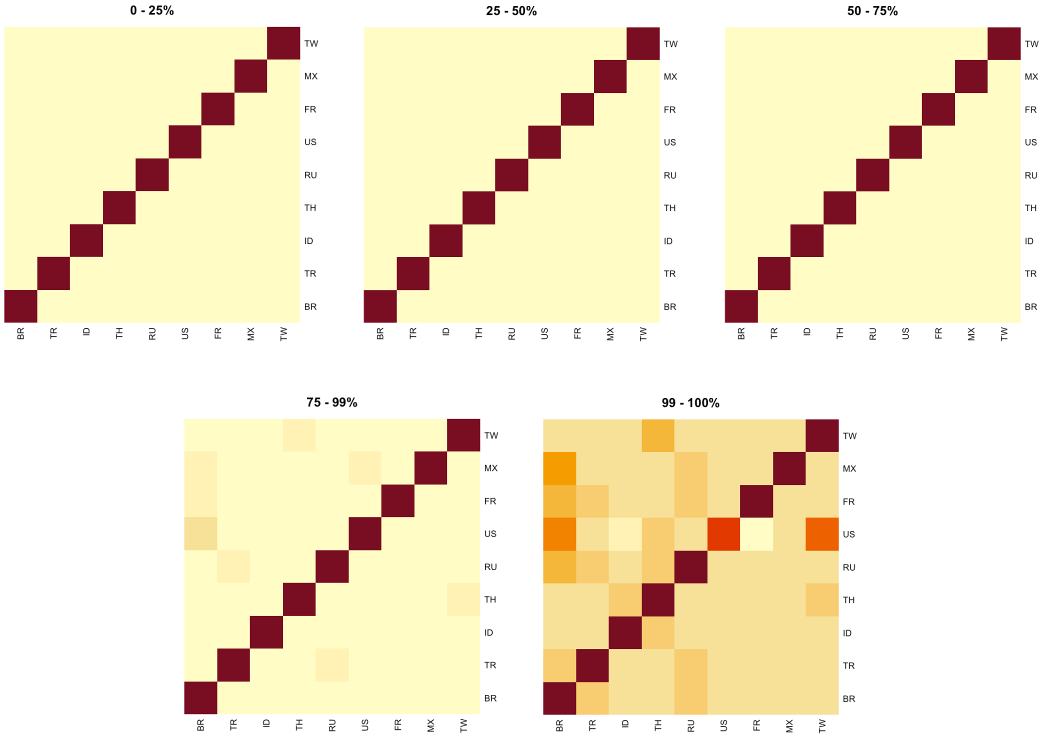
For celebrity type users, 77.80% users link their email address to the account, 40.99% users log in with a iOS device, 97.62% users set their profile visible, and 88.05% users have uploaded a photo.

We could compute the same descriptive statistics for the regular type users. For this group, 74.89% users link their email address to the account, 28.61% users log in with a iOS device, 97.85% users set their profile visible, and 70.21% users have uploaded a photo. It seems that regular type users prefer the Android platform, and they are less likely to upload photos.

3.4 Edges and Topology

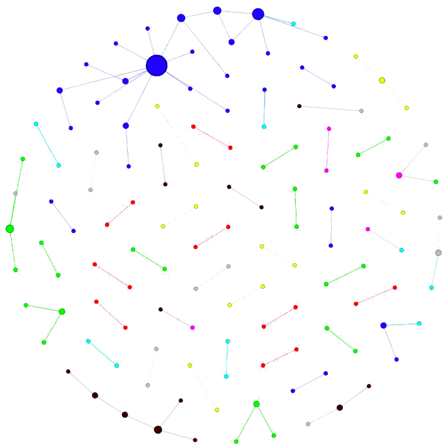
The proportions of top country users are quite stable across all degree quantiles. Therefore, a natural question that may rise would be whether top country users mainly connect with top country users as well. The resulting heatmaps in **Figure 2** below demonstrated that users in top countries build connections internally. Regardless of countries and degrees, 80% or 90% of the followers come from the same country.

Figure 2: Connection Heatmaps of 9 Top Countries at Different Quantiles



To understand the structure or characteristics of a social network, making a graph for its topology would be the most straightforward way. The following **Figure 3** plots a 5000-node stratified sample concerning the node degree which forms an almost disconnected graph.

Figure 3: Sparsity of the Degree of Normal Type Users

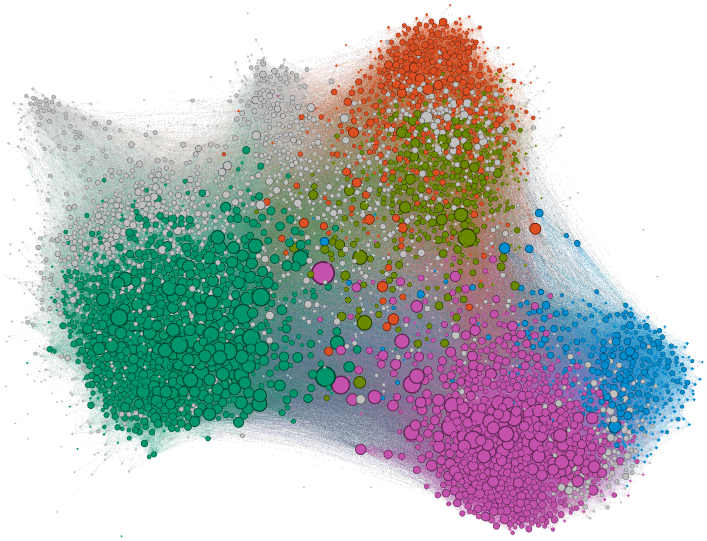


The topology of the sample could be seen as an estimate of the topology of the overall graph. It matches the expectation that the graph is quite sparse. Indeed, celebrity-type users own most of the connections and there are few connections among the regular type users.

4 Sampling Methods and Likelihoods

We compare the log-likelihood of five different sampling strategies on sub-setting the pairs in the first step of the algorithm. For instance, sampling according to informative pairs of nodes helps recover community structure, stratified sampling pays more attention to existing links so that it is more effective while applying to sparse networks, etc.

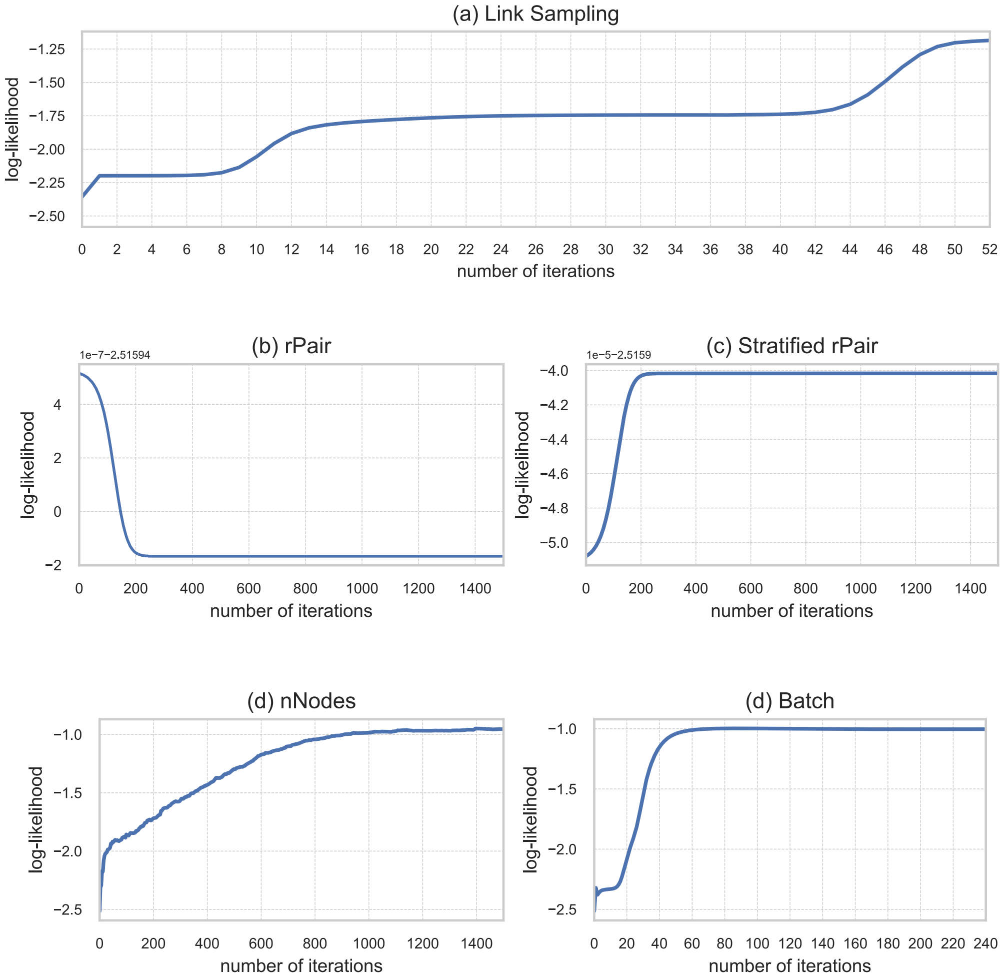
Figure 4: Network Topology of Community Assignments from Link Sampling



We use link sampling as our baseline method. Its 76 community assignments are illustrated in **Figure 4** with the five major communities specified in green (18.24%), pink (15.25%), red (11.24%), green (8.64%), and blue (7.5%). Compare to link sampling, the log-likelihood of batch sampling converges with the least iterations. However, the downside of batch sampling is that each iteration takes significantly more time than the baseline approach. These two approaches converge to -1.0 rather quickly within 50 iterations but at the efficiency angle, link sampling to some extend is still a more feasible approach.

As for random pair sampling, an instance of independent pair sampling means to sample node pairs uniformly at random. Unlike other sampling methods, the log-likelihood gradually decreases within the first 200 iterations and then stabilizes itself around -1.8. The abnormal pattern may be due to the aforementioned sparse nature of the network, therefore, the algorithms could struggle to find linked pairs. However, when we add a stratified component to the random pair sampling, it can be shown that the log-likelihood curve resumes increasing.

Figure 5: Log-likelihood of Different Sampling Methods



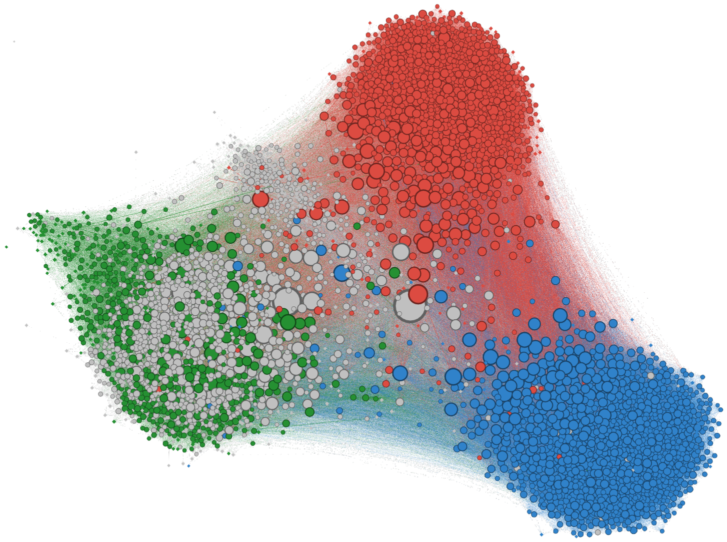
Stratified random pair sampling samples links independently but focuses more on observed links. All node pairs are divided into two strata: links and non-links. In each iteration, we either sample a mini-batch of links or sample a mini-batch of non-links. The log-likelihood for stratified random pair sampling has the normal rising pattern within the first 200 iterations, and then reaches its maximum of -4.1.

Instead of sampling the links, random node sampling is a method that concentrates on the local neighborhoods of the network. A set consists of all the pairs that involve one of the N nodes. At each iteration, we sample a set uniformly at random from all sets. Since each pair involves two nodes, each link appears in two sets. We maintain a correct stochastic optimization by re-weighting the terms corresponding to pairs in the sampled set. The log-likelihood for random node sampling converges to -1.0 around $1,000$ iterations.

5 Community Detection with Covariates

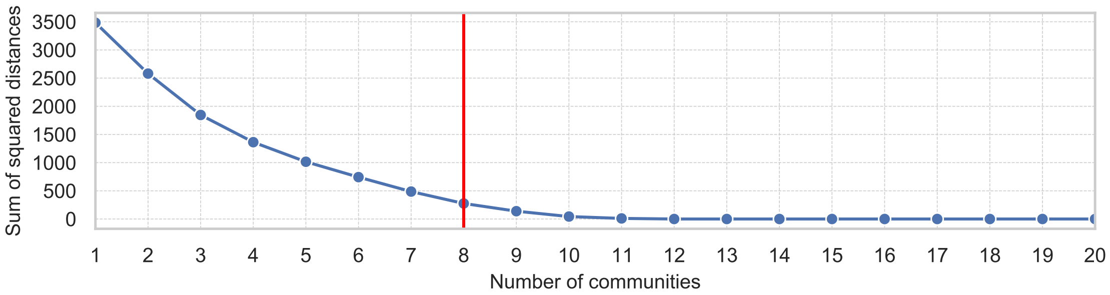
Previously, **Figure 4** illustrates the detected communities based on only nodes connection. In this section, we compare the community detection by using individual demographic characteristics.

Figure 6: Topology of Community Assignments Based on K-means Clustering



The demographic variables including user geographic longitude and latitude coordinates, the language spoke and nationality is used to project the nodes for K-means clustering. **Figure 6** demonstrates the result given 8 communities which are determined by **Figure 7** measuring the sum of squared distance under different numbers of communities. The three major communities are colored in red (27.35%), blue 26.75%, and green (14.27%). Similar results as **Figure 4** indicates that regardless of overlapping community, the demographic similarities between individuals act as a strong magnet that pulls people to form various communities.

Figure 7: K-means Sum of Squared Distances by Number of Communities



6 Contributions

We discussed the paper together and split all tasks evenly to make sure all the codes are working. In particular, Hanqiao provided descriptive statistics about the data. Jieyu ran the code provided in [1] using various sampling methods and plotted the log-likelihood of the holdout set of the graph sampled. Hao-Che visualized the community outcome using Gephi and ran a K-means classifier using the covariates data for comparisons.

References

[1] Prem K. Gopalan and David M. Blei. Efficient discovery of overlapping communities in massive networks. *Proceedings of the National Academy of Sciences of the United States of America*, 110(36):14534–14539, September 2013.

[2] Prem Gopalan, David Mimno, Sean M. Gerrish, Michael J. Freedman, and David M. Blei. Scalable inference of overlapping communities. *NIPS*, 2012.

[3] Tom A.B. Snijders and Krzysztof Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14:75–100, 1997.

[4] Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, May 2007.