EE 232E Project 4
IMDb Mining

Hengjie Yang, Sheng Chang, Wandi Cui, and Tianyi Liu

June 3, 2018

1 A brief tutorial on how to use this template

Please remove the tutorial section in the final manuscript by commenting, i.e. %(*something*)

1.1 Figures

Figure insertion is shown in Fig 1.

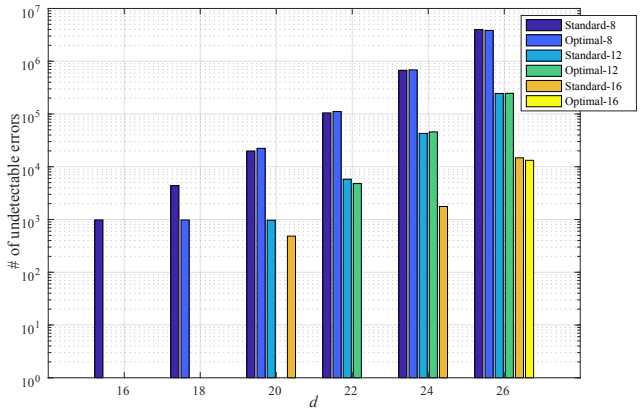

Figure 1: An example of figure insertion

1.2 Equations

An example of equations is given as follows.

Theorem 1. *Let a, b, c denote the sides of a triangle, respectively. If $a \perp b$, the pythagoras theorem is given as follows.*

$$c^2 = a^2 + b^2 \tag{1}$$

1.3 Tables

An example of tables is shown in Table 1.

Table 1: Standard CRC Codes versus Optimal CRC Codes for Convolutional Code $G = (561\ 753)$ with $n = 504$ Bits

Name	Gen. Poly.	Undetected Error Distance Spectrum						
		d	16	18	20	22	24	26
Standard-8	0x19B		983	4387	19909	105000	672724	3972970
Optimal-8	0x19D		0	979	22349	111304	686314	3830340
Standard-12	0x180F		0	0	969	5815	42893	245211
Optimal-12	0x108B		0	0	0	4793	45795	246729
Standard-16	0x11021		0	0	484	0	1765	14752
Optimal-16	0x1F8FD		0	0	0	0	0	13240

1.4 Actor rankings

We aimed to find to find the top 10 actor/actress in the network using the google's pagerank algorithm. Those information of the top 10 actor/actress is shown in Table 2, including the name, the number of movies and the in-degree of each of the actor/actress in the top 10 list.

We can see from the result that it does not have any of the actor/actress listed in the previous section. In general, the more movie they took part in, the high pagerank they may had, because that means they had more changes to cooperate with other actor/actress and it's obvious that they may have higher degree in the network. After googling it, we found that most people int the top 10 are actually voice actors. That's why they can take part in hundreds of movies and that also explains why those famous previous actor/actress are not included in the

Table 2: Top 10 highest pagerank score actor/actress

Name	the Number of Movies	In-degree
Flowers, Bess	828	7537
Tatasciore, Fred	355	3954
Harris, Sam (II)	600	6960
Blum, Steve (IX)	373	3316
Miller, Harold (I)	561	6587
Jeremy, Ron	637	3177
Phelps, Lee (I)	647	5563
Lowenthal, Yuri	318	2662
Downes, Robin Atkin	267	2953
O'Connor, Frank (I)	623	5502

top 10. Even though those movie super stars acted so many movies, it's very common that they still act less than those voice actors.

What's more, the same information of the actor/actress listed in the previous section is shown in Table 3.

Table 3: The same information table for previous actor/actress

Name	the Number of Movies	In-degree
Tom Cruise	63	1651
Emma Watson (II)	25	453
George Clooney	67	1573
Tom Hanks	80	2064
Dwayne Johnson (I)	78	1357
Johnny Depp	98	2144
Will Smith (I)	49	1319
Meryl Streep	97	1594
Leonardo DiCaprio	49	1301
Brad Pitt	71	1739

2 Movie Network

2.1 Undirected movie network creation

We create a weighted undirected movie network. And the degree distribution of the movie network is shown in Figure 2. We can see from the result that most movies have a degree between 500 and 1000, also there're only a few movies that have very large or very small degrees. The result is not super surprising since it is very common that lots of movies share same popular movie starts for the box office.

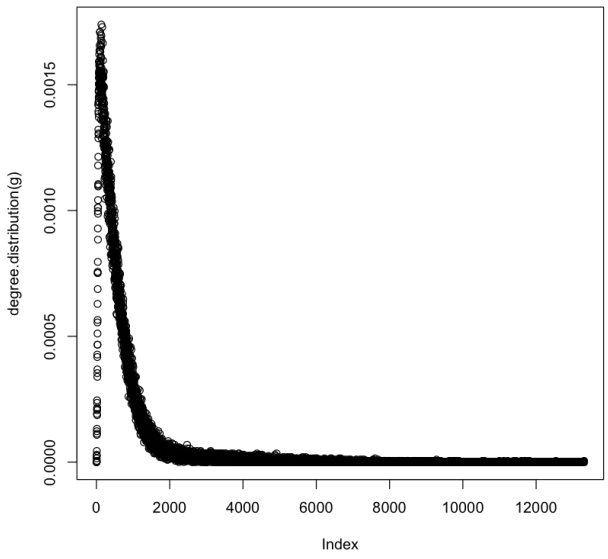

Figure 2: Degree distribution of the movie network

2.2 Communities in the movie network

By detecting communities with Fast Greedy algorithm for our movie network, we got 28 communities. And for Quesition 7, we just picked the first 10 communities to plot the distribution of the genres of the movies in each community.

NOTICE: Some of the movies' genre information is missing in the given dataset, so we marked them as "NAN"; however, only Question 7, we considered them in our plots.

The 10 plots are shown in Figure 3 to Figure 12.

Based on simple frequency counts, the most dominant genre of each community we got is shown in Table 4. As we can see, the genre "Drama" is the dominant genre of 11 communities among 28 communities. Thus, genre "Drama" tends to be the most genre among all communities.

Why???

Based on the measurement of modified score, the dominant genre of each community changes a lot, and the result is shown in Table 5. It illustrates that most of the dominant genres we got according to the modified scores are different from the genres we got by simply couting the frequency of genre within the community. Interestingly, there do exist some communities hold the same dominant genre and all of them have comparably fewer movies than other communities. To be sepecific, the number of movies in these communities is in the range of (10,20). In addition, the frequency of the dominant genre in these communities is far more higer than other genres. For instance, we also plotted community with ID=24 and ID=25, shown in Figure 13 and 14. In fact, this is reasonable because in these kinds of communities, $p(i)$ of the dominant genre(the genre has the highest frequency) is much bigger than other genres within the community, which can play a dominant role of the score function. Moreover, some of the genres in these kinds of communities have only one occurency, which makes $c(i)$ equals to 1 and then $ln(c(i))$ equals to 0, so that the score will be 0. On the contrary, for those communities whose dominant genre differ from different measurement, they generally include large number of movies, and may have genres holding the similar frequency wihthin the communiy. For instance, we could observe the genre distribution of community with ID=6 (Figure 8), the dominant genre changes from "Drama"(the most frequent genre)

to "Comedy"(the third frequent genre). This is easy to understand since the score function consider the fraction of genre within community and also in whole dataset, which is equvalent to multiply some coefficients to the exact frequency of the genre, this somehow makes the score bigger or smaller comparing to the exact frequency. Therefore, the dominant genre of these kinds of community changes.

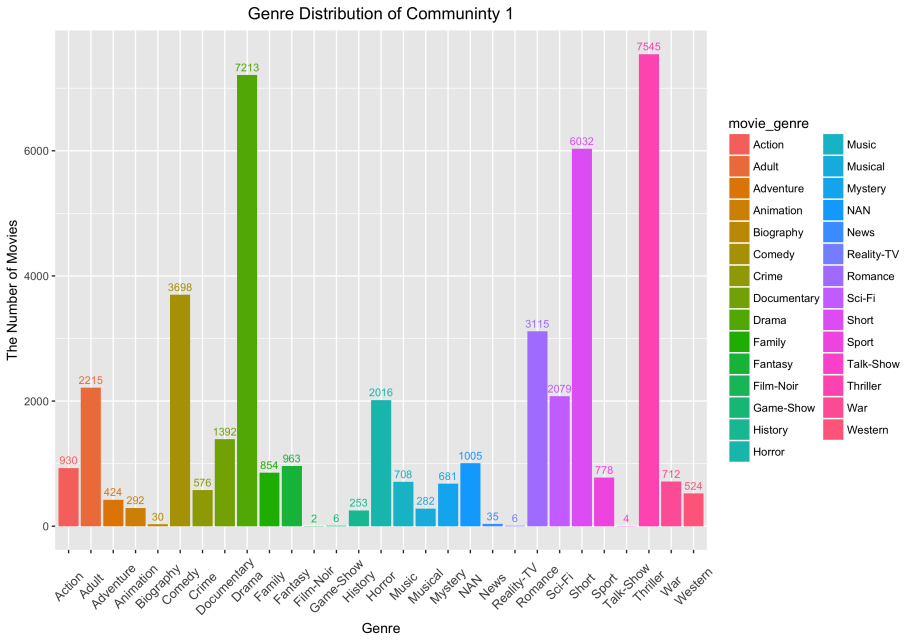

Figure 3: Genre Distribution of Community 1

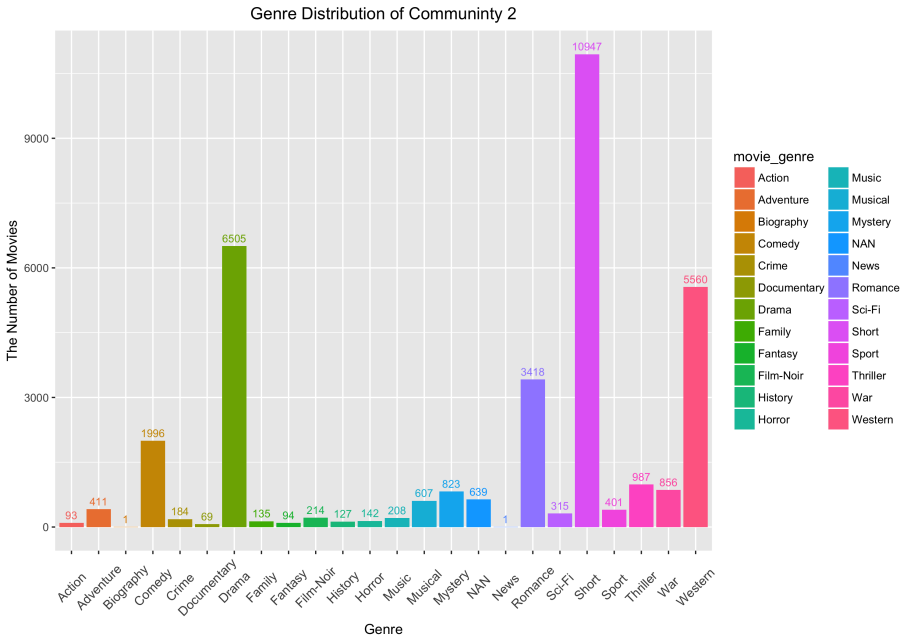

Figure 4: Genre Distribution of Community 2

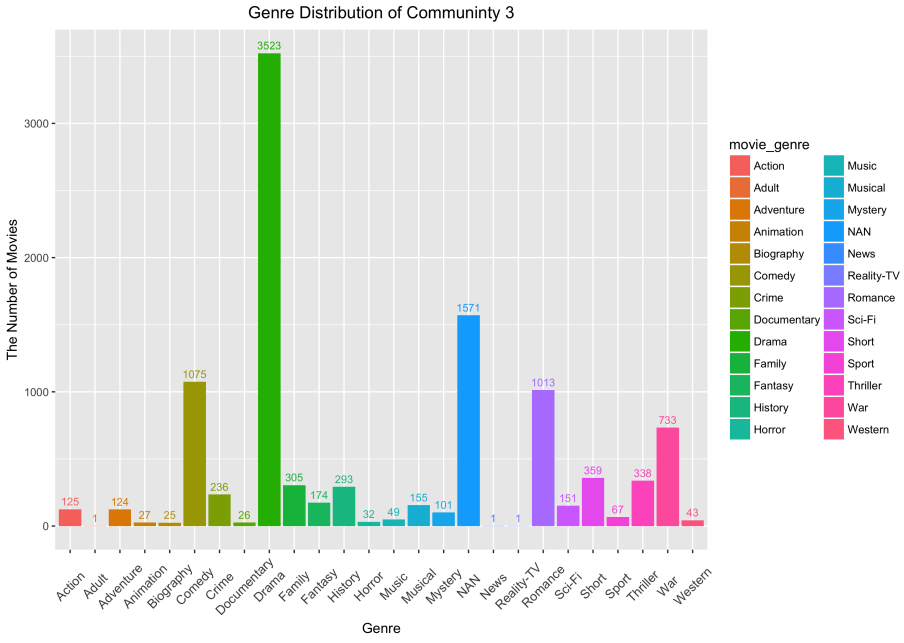

Figure 5: Genre Distribution of Community 3

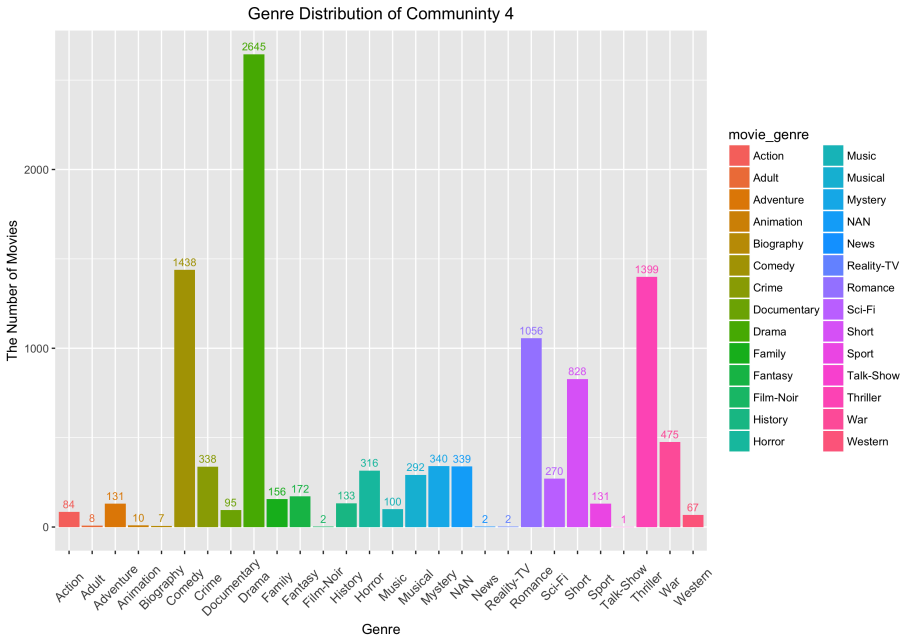

Figure 6: Genre Distribution of Community 4

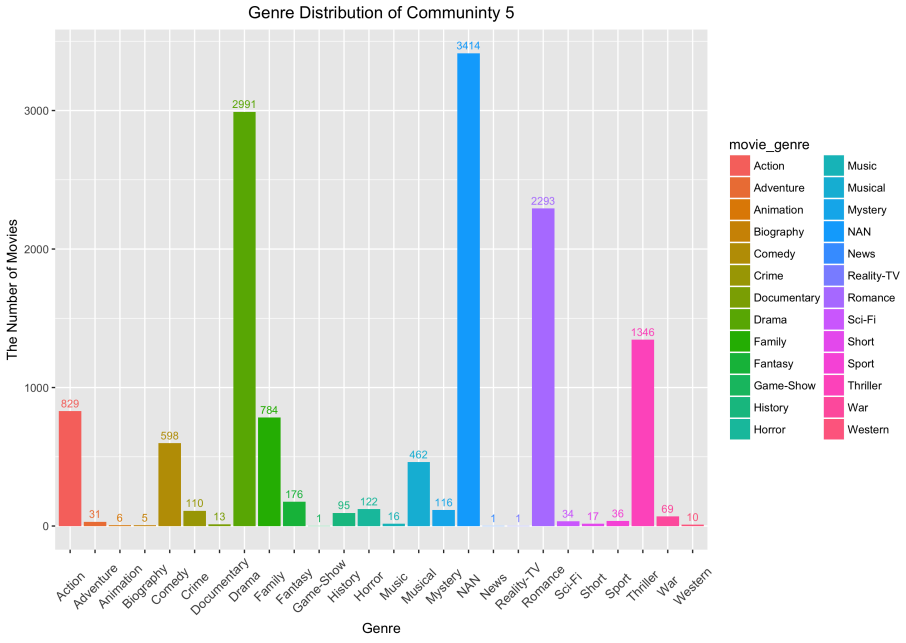

Figure 7: Genre Distribution of Community 5

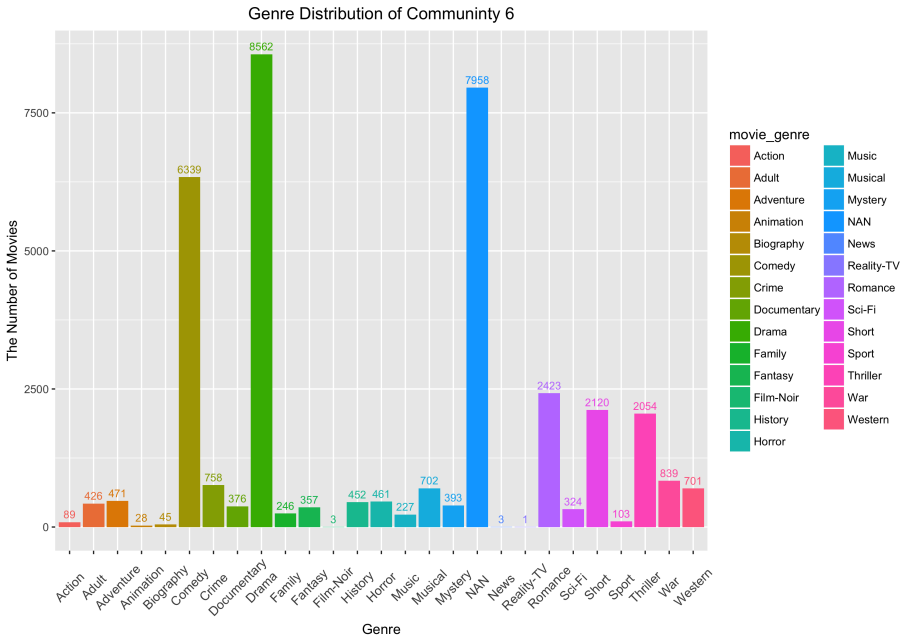

Figure 8: Genre Distribution of Community 6

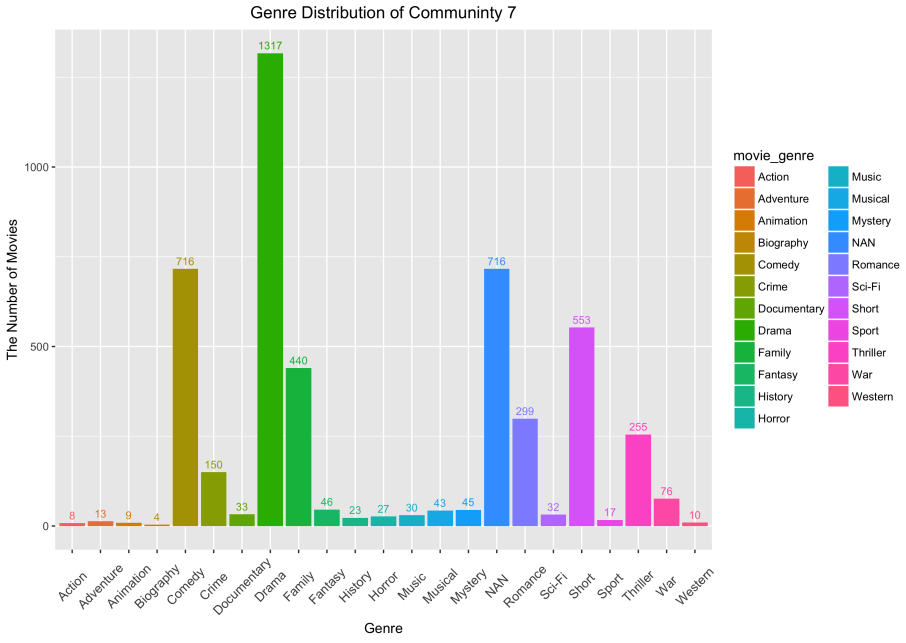

Figure 9: Genre Distribution of Community 7

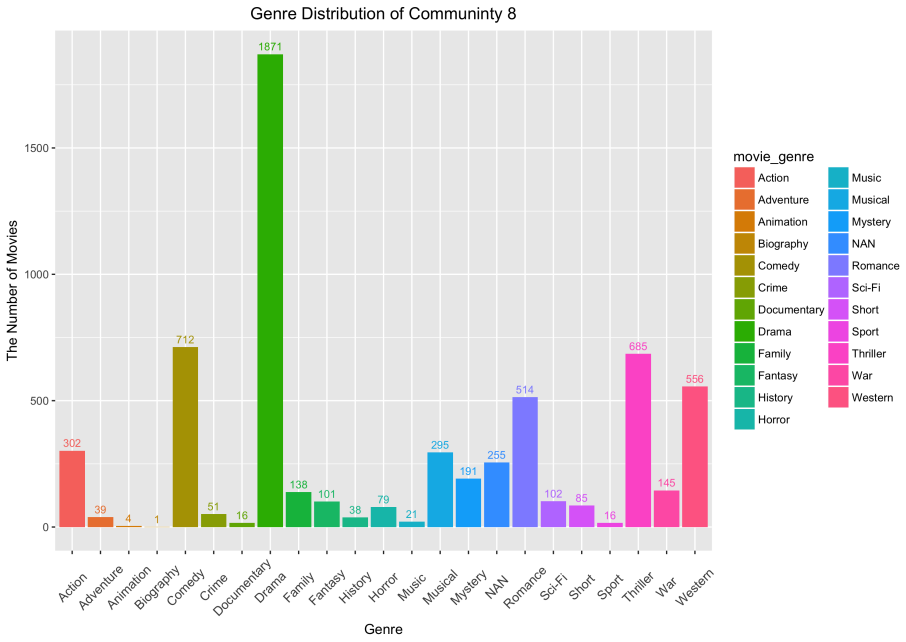

Figure 10: Genre Distribution of Community 8

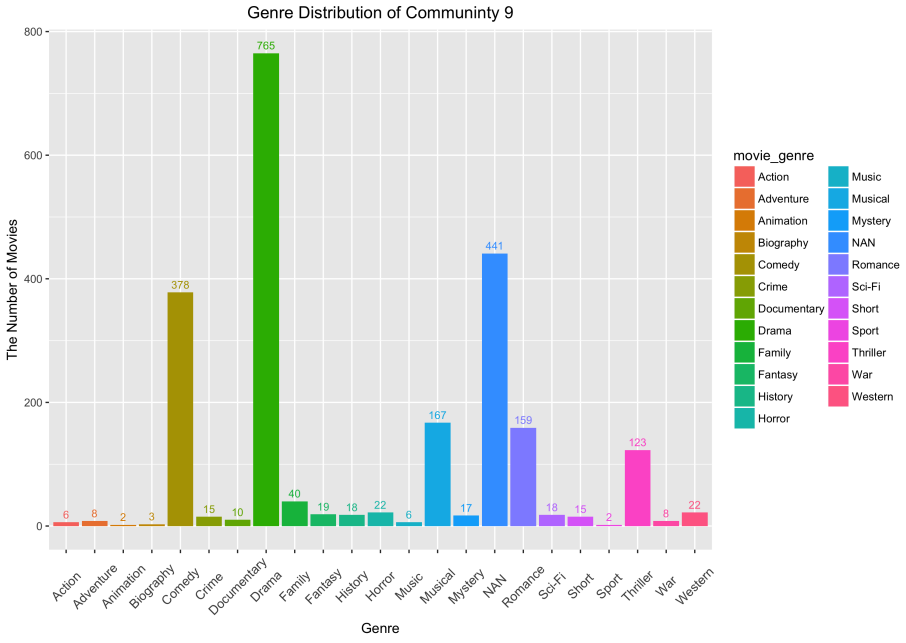

Figure 11: Genre Distribution of Community 9

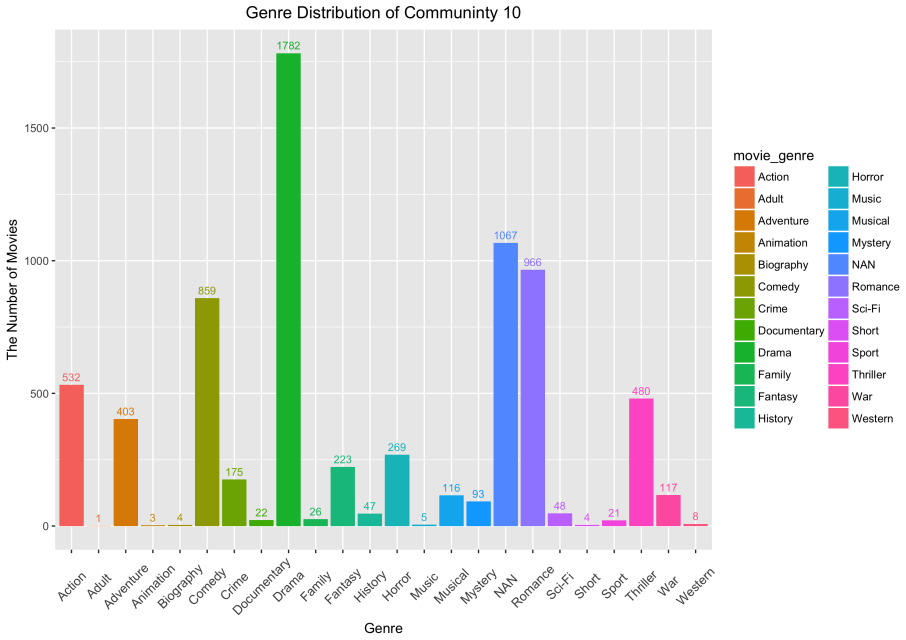

Figure 12: Genre Distribution of Community 10

Table 4: The dominant genre of each community(according to frequency count)

Community ID	**the dominant genre**
1	Thriller
2	Short
3	Drama
4	Drama
5	Drama
6	Drama
7	Drama
8	Drama
9	Drama
10	Drama
11	Drama
12	Drama
13	Drama
14	Drama
15	Drama
16	Drama
17	Drama
18	Drama
19	Drama
20	Drama
21	Drama
22	Drama
23	Drama
24	Adult
25	Thriller
26	Short
27	Short
28	Short

Table 5: The dominant genre of each community(according to modified score)

Community ID	**the dominant genre**
1	Adult
2	Film-Noir
3	War
4	Crime
5	Family
6	Comedy
7	Family
8	Musical
9	Musical
10	Adventure
11	Family
12	Romance
13	War
14	Adventure
15	Comedy
16	Musical
17	Action
18	Drama
19	Fantasy
20	Comedy
21	Action
22	Romance
23	Short
24	Adult
25	Thriller
26	Short
27	Short
28	Short

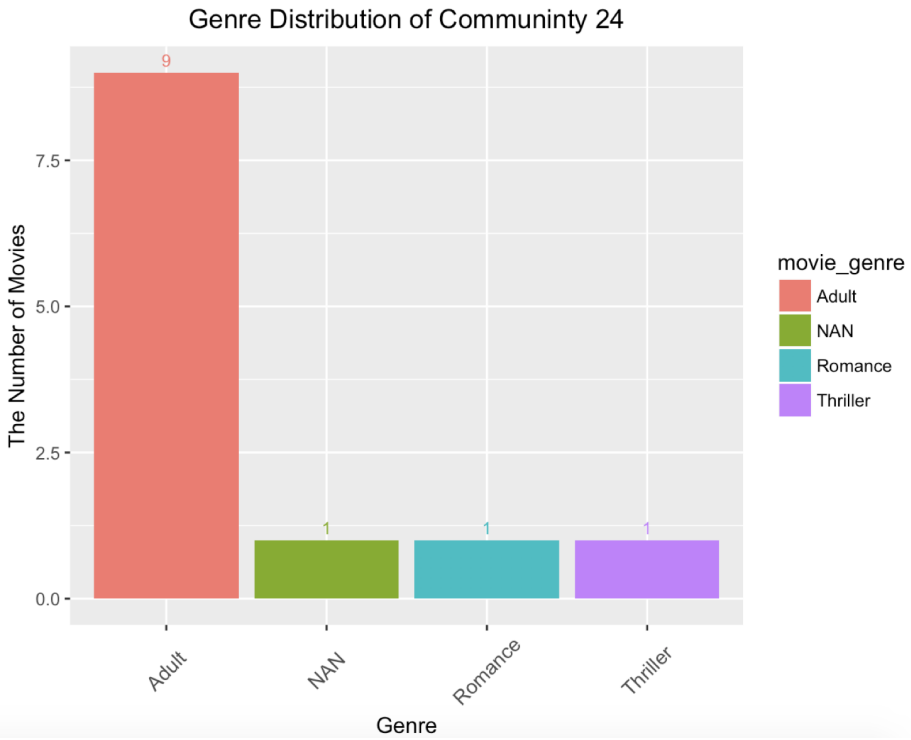

Figure 13: Genre Distribution of Community 24

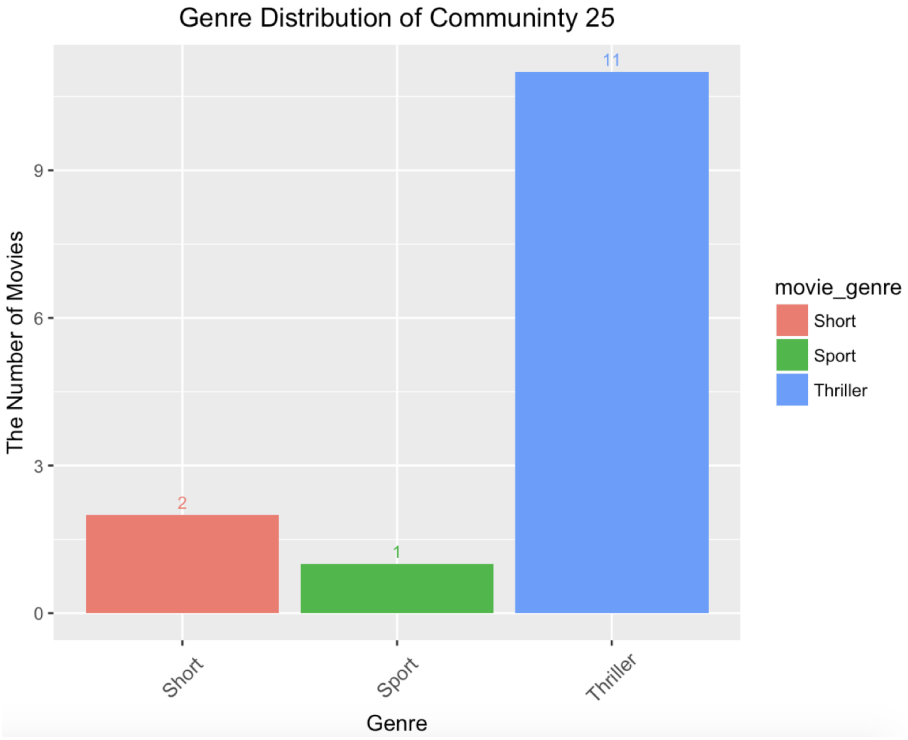

Figure 14: Genre Distribution of Community 25