NUS National University of Singapore
School of Computing

Team Number: CLUSTER2-10

BOOKS TO MOVIES

Website: https://yee172.github.io/Books2Movies/

Group Members

Tian Fengrui, XJTU

Peng Jiawei, ZJU

Ye Qihao, SUSTech

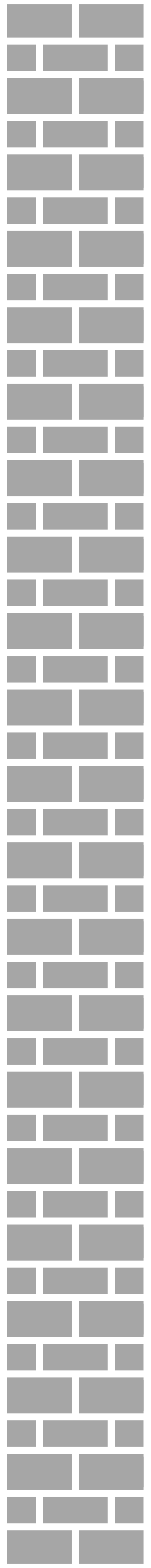
Liu Yihong, SCU

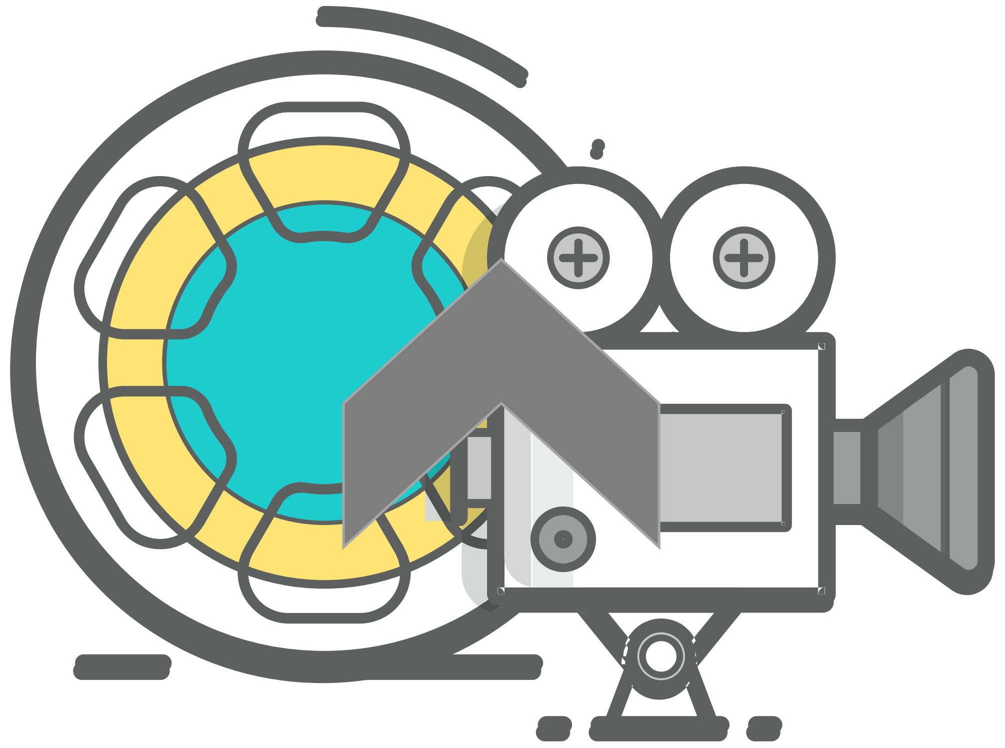
CONTENTS

Introduction

How to recommend Movies based on books

Introduction



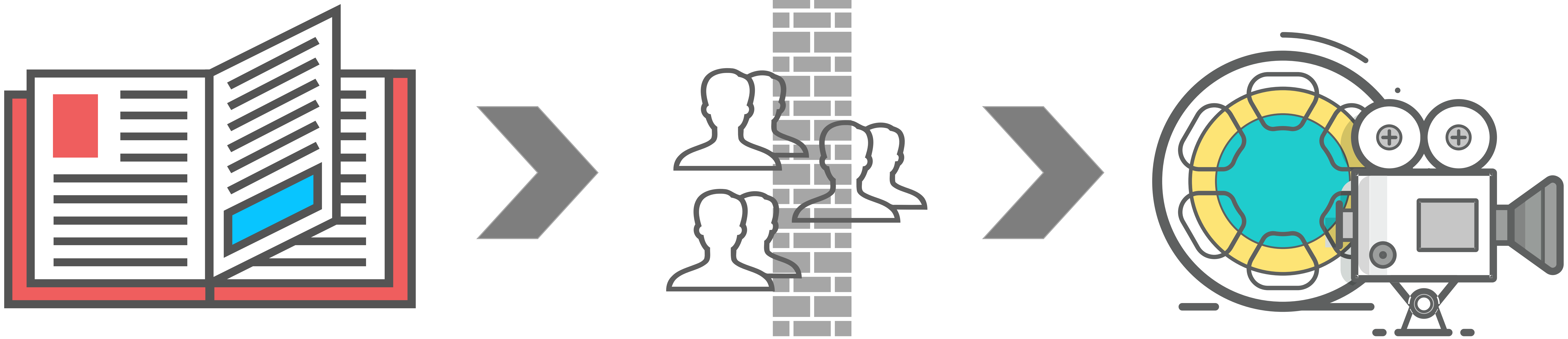


Introduction

School of Computing

First Graph

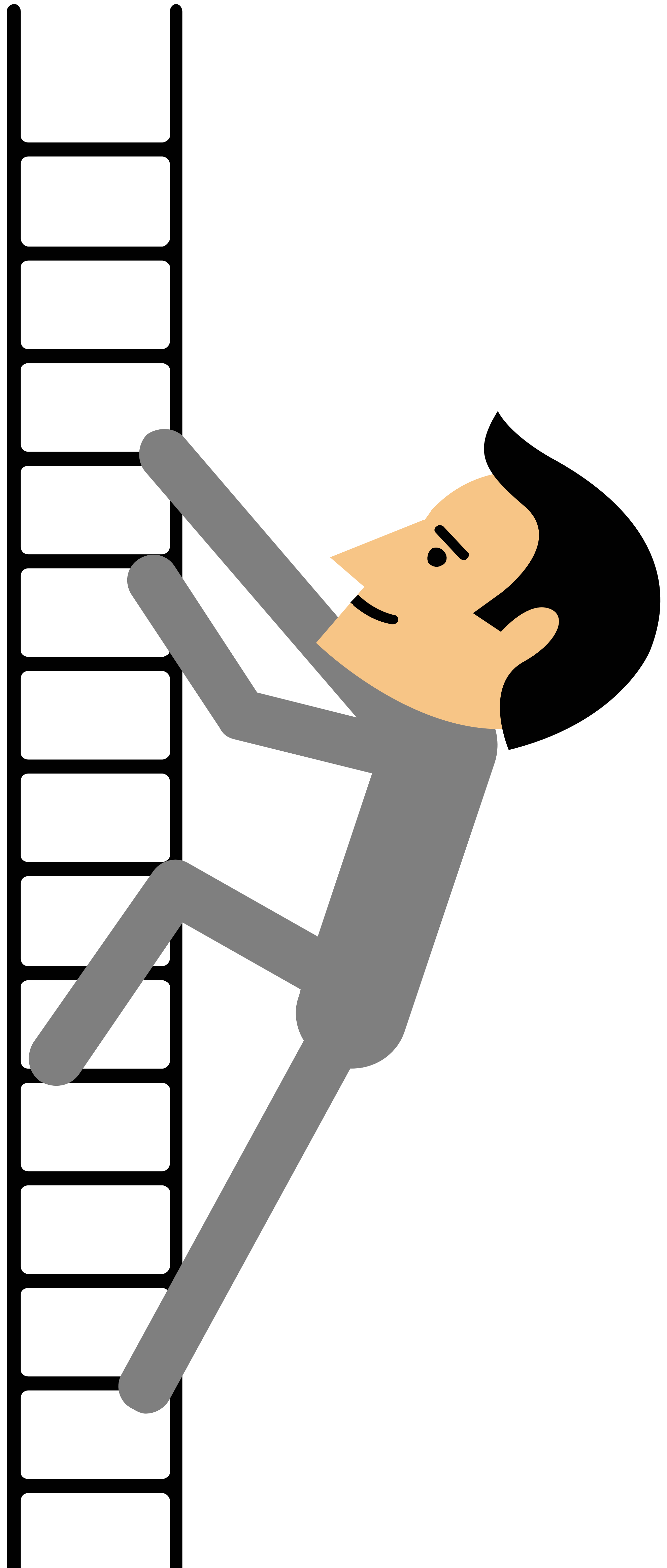
Second Graph



WORKFLOW

New dataset, new algorithms, new graphs.

Data Collecting



01 Different Books

Using different tags to separate books. 7 kinds of books in total.

02 Users' Informations

Collecting user' info in different book groups.

03 Users' Preference in Movies

Collecting each user's movies he/she has watched.

04 Movies' Informations

Getting names, directors, scriptwriters, actors, types, etc.

Data Processing-Step 1

This time to make the dataset more representative, we decide to use Chinese library classification

Big Tag	Sub Tag	Book
文学	诗词	苏东坡文选……
	散文	xx散文集……
	随笔	鲁迅杂文…
……		
科普	网络	……
	编程	……
	算法	……

For every subtag here we find top15 books here. This may ensure the result of first graph better, and in a consequence making our findings better.

Data Processing-Step 1

After counting out those who read all kinds of books equally frequent, then select the top 125 active users to build the graph.

B1 = Literature

B2 = Popular

B3 = Culture

B4 = Philosophy

B5 = Military

B6 = Economic

B7 = Art

B8 = Science

4 ★
5 ★

User_id	Book_id	Category
1	1	B1

↓

User_id	B1	B2	……	tag
1	1	0		Not decided

Data Processing-Step 1

NUS National University of Singapore
School of Computing

When the calculation is done, we can generate their tags, which is the top3 kinds of book they prefer.

User_id	B1	B2	B3	B4	……	Big tag
0	11	22	1	2		B2|B1|B6
1	17	0	3	12		B7|B5|B4
……						

Data Processing-Step 2

After counting out those who read all kinds of books equally frequent, we select the top 125 active users to build the graph.

The first graph are 105 nodes and 6 clusters. We will analysis the clusters.

The number of people who watched this movie in each cluster represents how it is liked

Figure out the movies that these users have seen.

Data Processing-Step 2

School of Computing

Every cluster in Graph1 may have some meanings, so we use them as the element of the movies' vector.

C1 = Economic

C2 = Popular

C3 = Philosophy

C4 = Science

C5 = Philosophy&Science

C6 = Literature

Cluster	Movie_id	Number of like
1	1	10

↓

Movie_id	C1	C2	……
1	10	0	

Data Processing-Step 2

School of Computing

When the calculation is done, we can generate their tags according to how they are liked by the clusters, and compare them with their original tags

	Movie_id	cluster1	cluster2	cluster 3	……	OriginTag	Cluster_tag
Inception	0	12	7	15	……	Science Fiction | Suspicion | Adventure	C6
Life of Pi	1	3	5	12	……	Fantasy | Adventure	C6
	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Dismat Calculation

School of Computing

Cosine similarity:

$$\text{similarity}(u,v) = \frac{|\min(\mathrm{N}(u),\mathrm{N}(v))|}{\sqrt{|N(u)||N(v)|}}$$

Jaccard similarity:

$$\text{similarity}(u,v) = \frac{|\min(\mathrm{N}(u),\mathrm{N}(v))|}{|\max(\mathrm{N}(u),\mathrm{N}(v))|}$$

UserCF-IIF:

$$\text{similarity}(u,v) = \frac{\sum_{i\in\text{ITEM}} \frac{\min(u[i],v[i])}{\log(1+|\mathrm{N}(i)|)}}{\sqrt{|N(u)||N(v)|}}$$

similarity to distance:

$$\text{distance}(u,v) = 1 - \text{similarity}(u,v), \text{similarity} \in [0,1]$$

$$\text{distance}(u,v) = \frac{1}{1+\text{similarity}(u,v)}, \text{similarity} \in [0,+\infty)$$

normalization:

$$x' = \frac{x-\min(X)}{\max(X)-\min(X)}, x' \in [0,1]$$

$$x' = \frac{x-\mu}{\max(X)-\min(X)}, x' \in (-\infty,+\infty)$$

$$x' = \log(1+x), x' \in [0,+\infty)$$

$$x' = \log(x), x' \in (-\infty,+\infty)$$

$$x' = \frac{2\arctan(x)}{\pi}, x' \in [0,1)$$

$$x' = \frac{x-\mu}{\sigma}, x' \in (-\infty,+\infty)$$

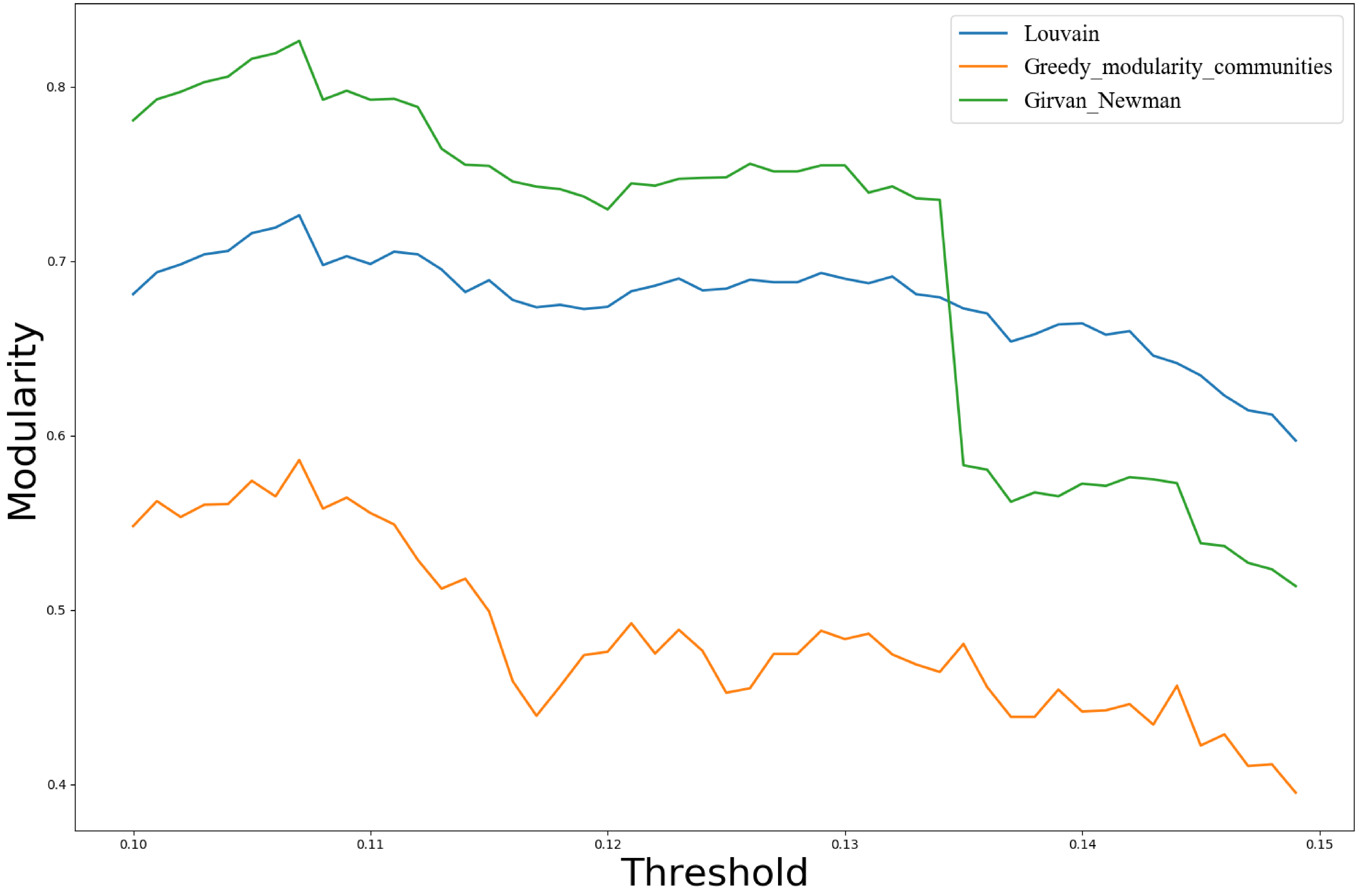
$$x' = x-\mu, x' \in (-\infty,+\infty)$$

Algorithm Application

We want to choose the best algorithm and a better threshold,

so we compare the algorithm and the threshold by using this plot

Graph1

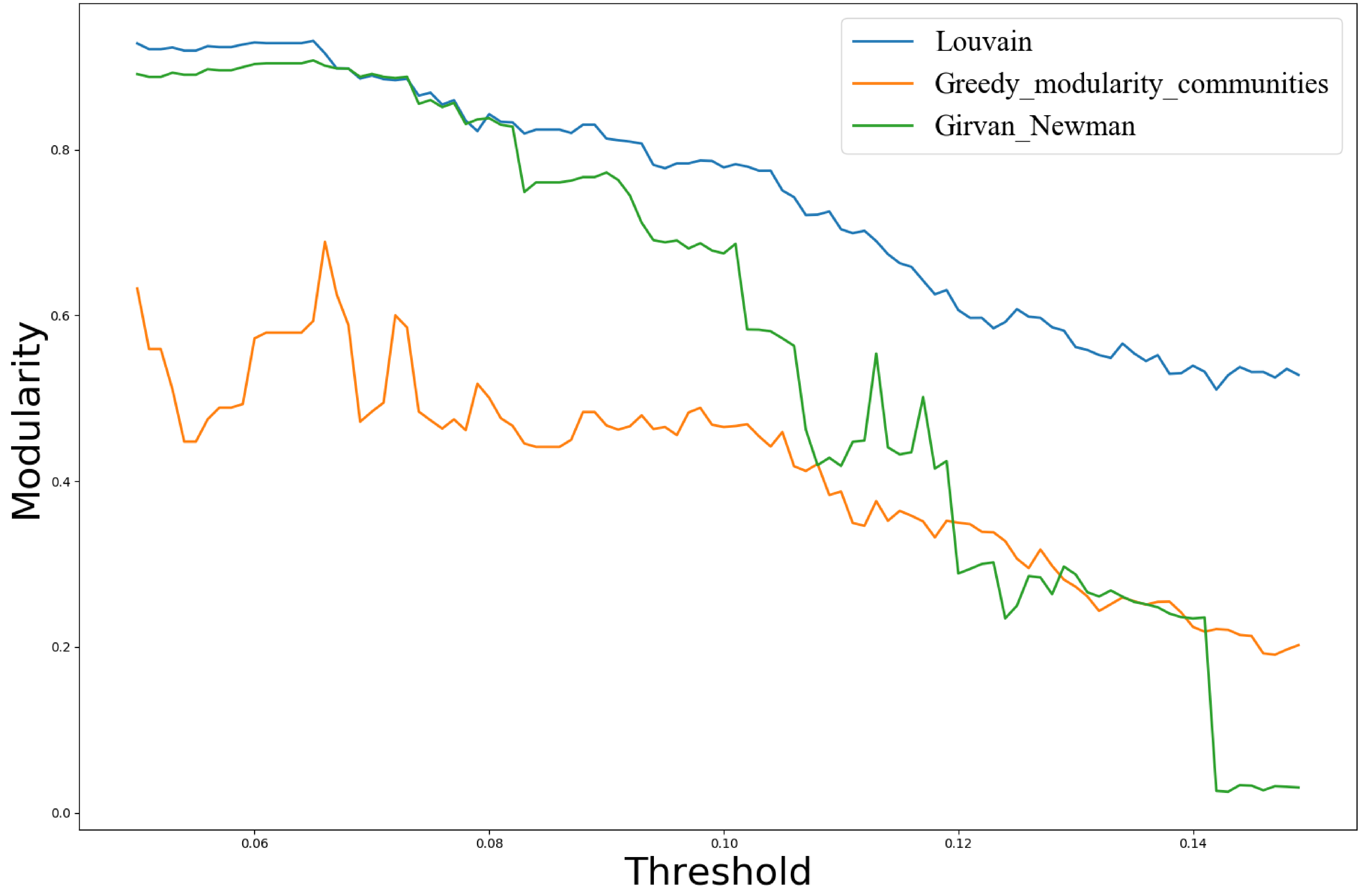


Girvan Newman works much better before the threshold goes beyond 0.135

Algorithm Application

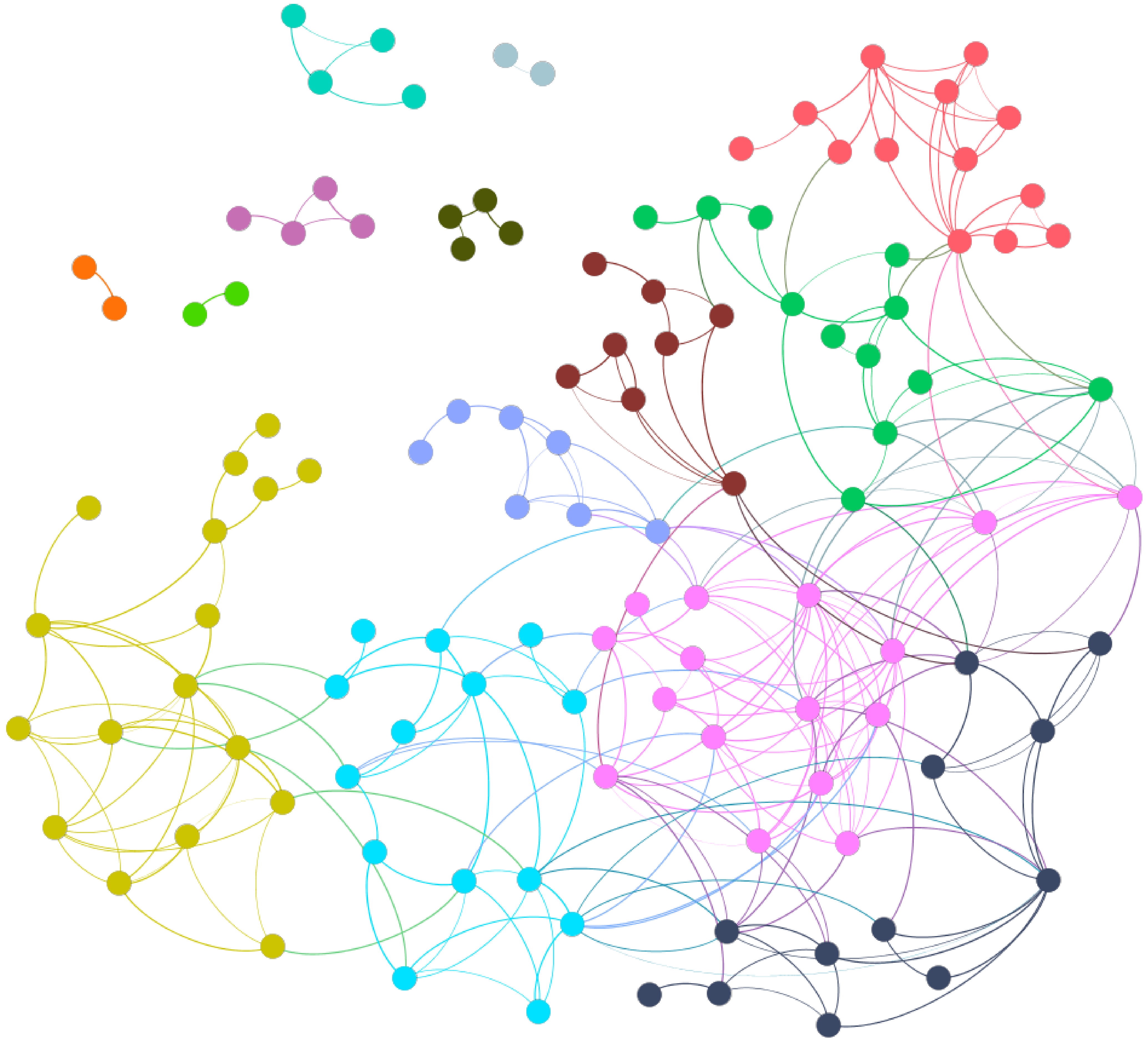
We assume things may change when it comes to the second graph, and it really happened

Graph2



This time, Lonvain always works better

Graph

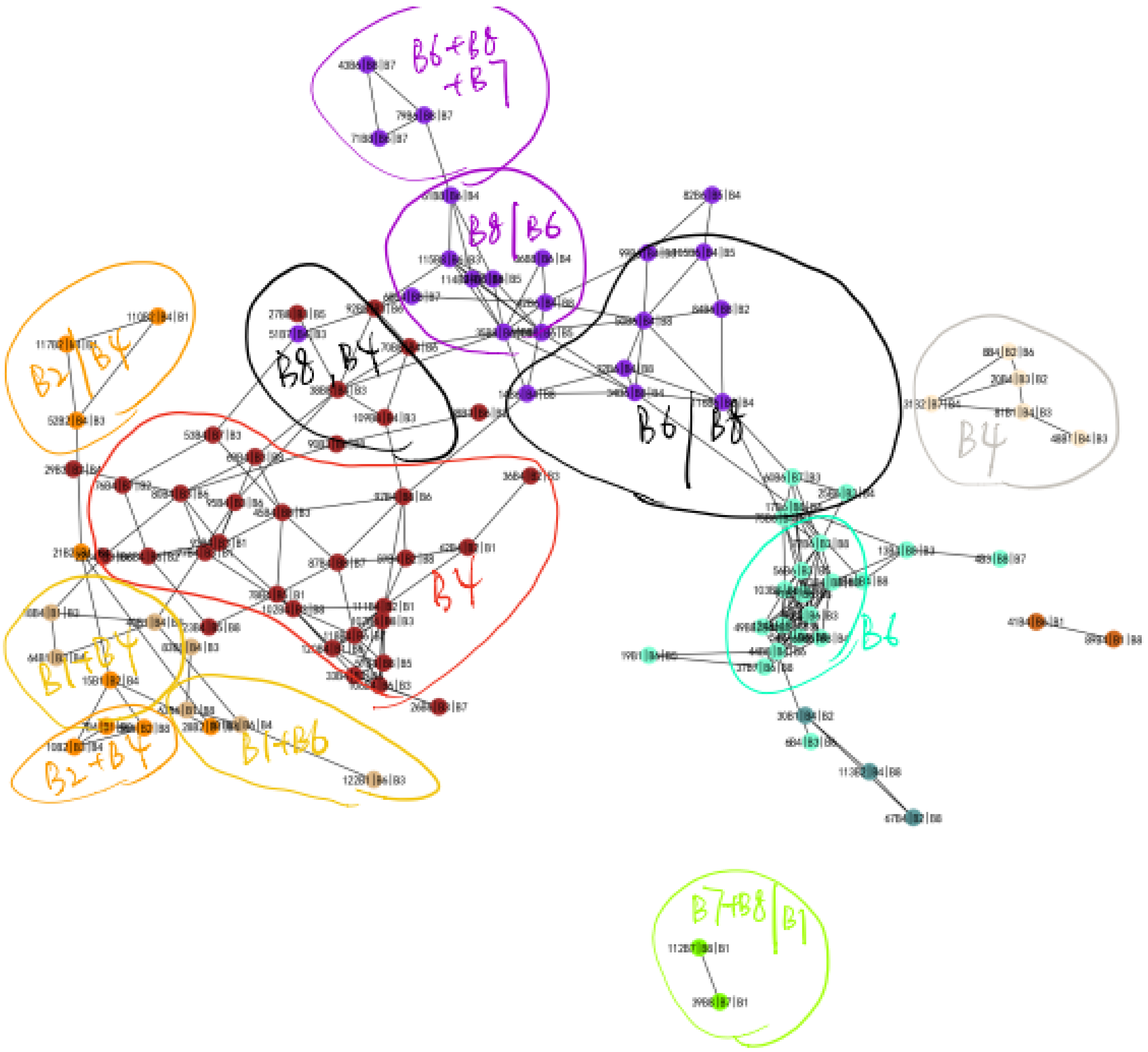


- Family + Sci-Fi
- Comedy + Romance C2+C3
- High tense but not irritating C2
- Classic
- Suspense C4
- Drama C6
- Popular
- Exciting Movie C4
- Sci-Fi C5/C1
- Popular Adventure
- Popular comedy
- Popular High rating
- Life Philosophy Sense of Time

Interesting Results

What we know we know, what we know we don’t know
What we don’t know we know, what we don’t know we don’t know

Preview: Graph 1



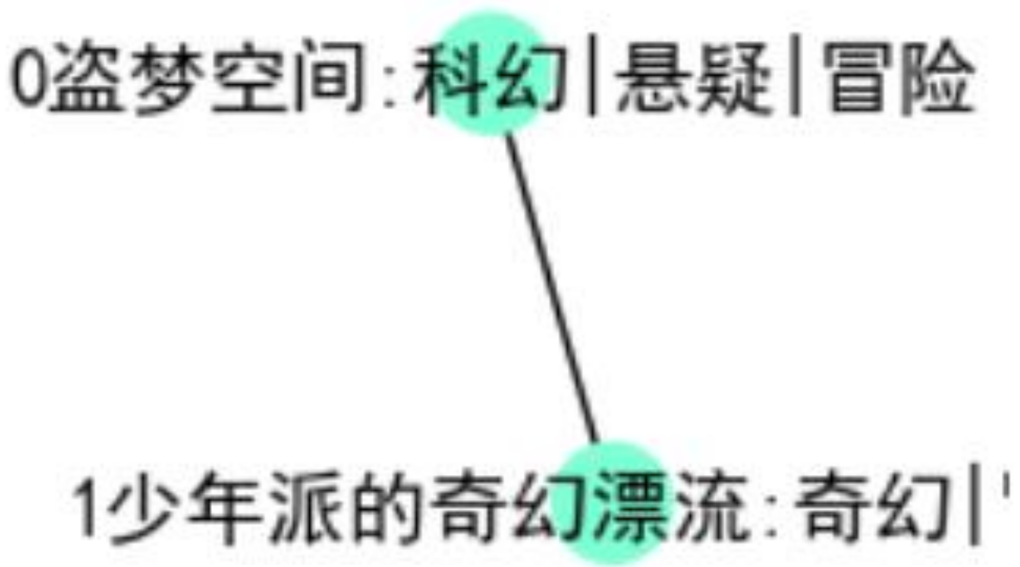
Though the first graph only serves as a bridge from book to movie but it has some meaning.

Every community here share the same preference.
The boundary areas have a mixed preference of two kind of books.

This means the result of first graph is valid.

Summary

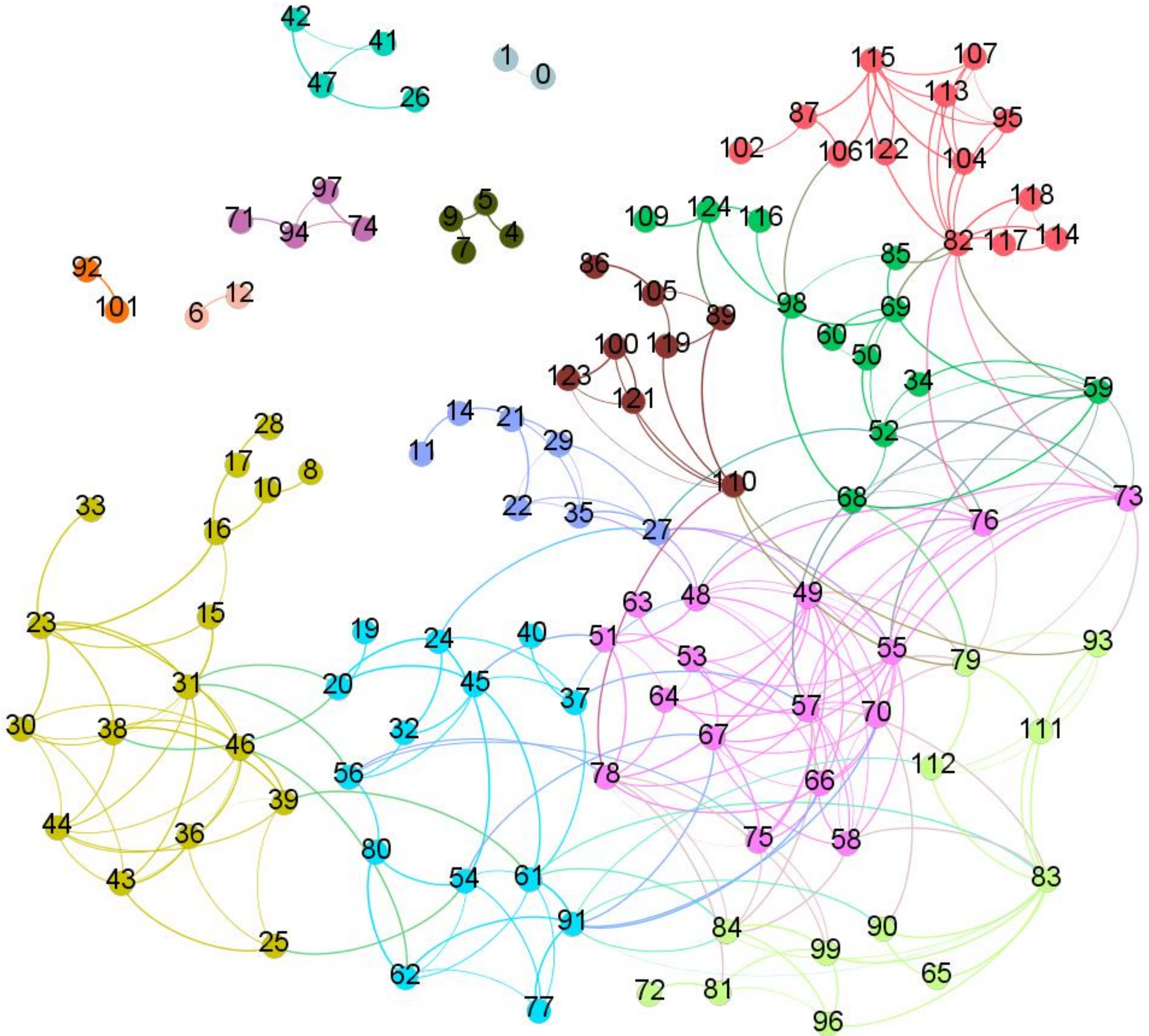
In general, many of the edges in the graph are predictable, but there are also many edges that we have not predicted. For example, the edge between ***Inception*** and ***Life of Pi*** are predictable, but the edge between *Youth* and *A Chinese Odyssey Part Two: Cinderella* is not expected.



What We Already Know

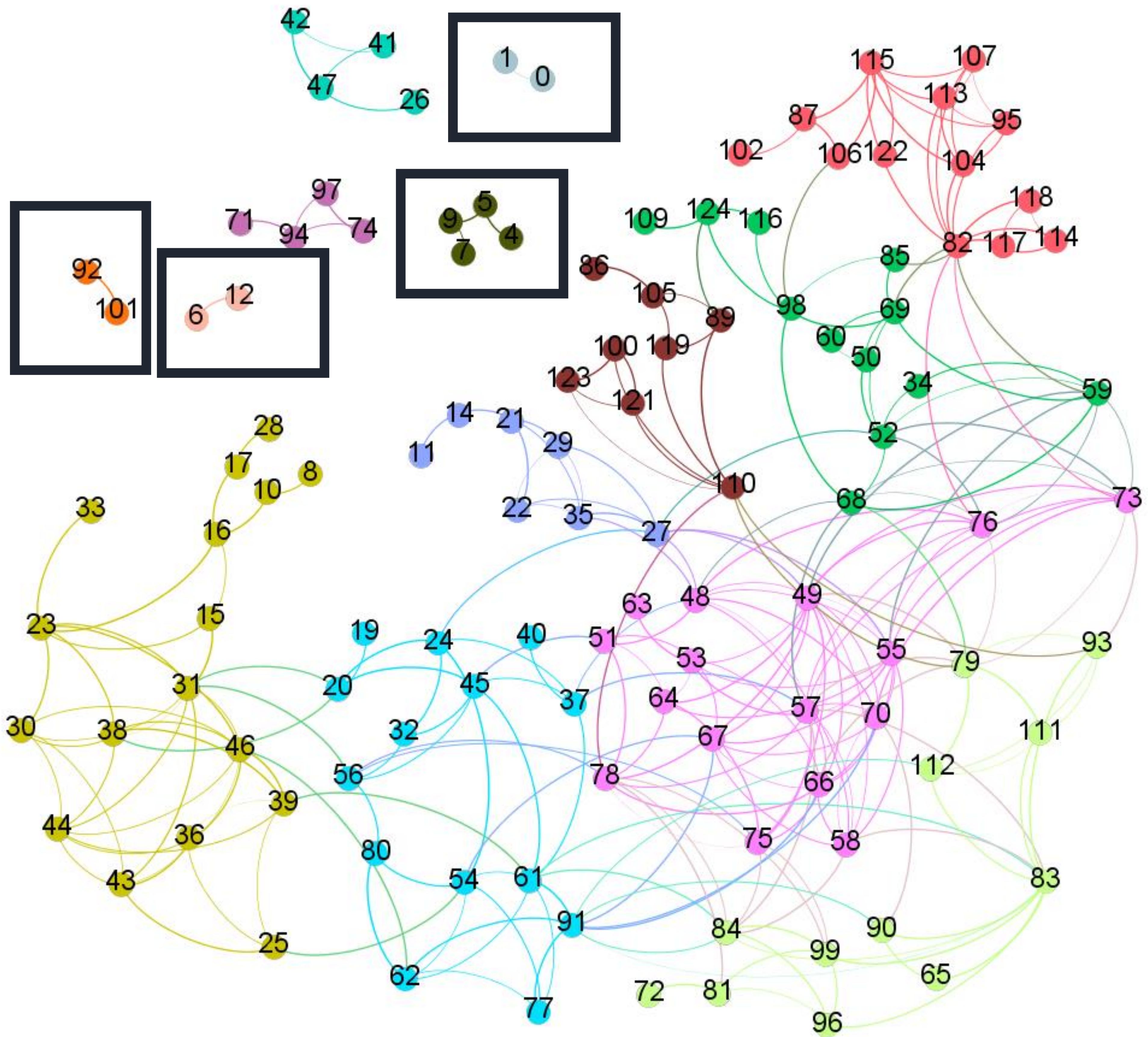
NUS National University of Singapore
School of Computing

Of the most film types, there are always one or two special ones that are especially popular among people

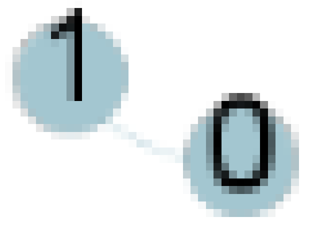


What We Already Know

Of the most film types, there are always one or two special ones that are especially popular among people



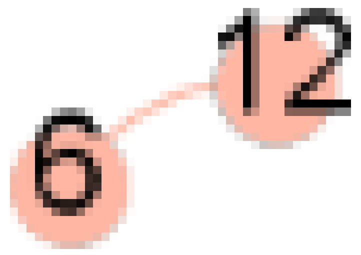
What We Already Know



Inception

Life of Pi

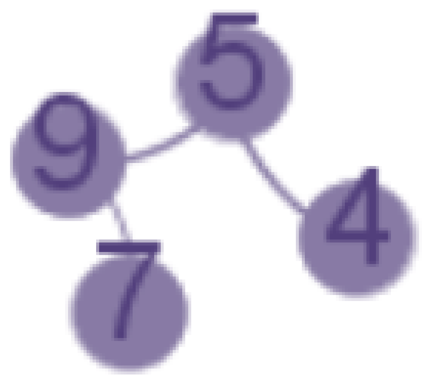
Most popular sci-fi



Let The Bullets Fly

Zootopia

Most popular comedy



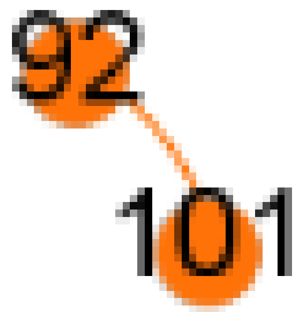
The Shawshank

Farewell My Concubine

Forrest Gump

3 idiots

TOP10 high-rating movies



The Butterfly Effect

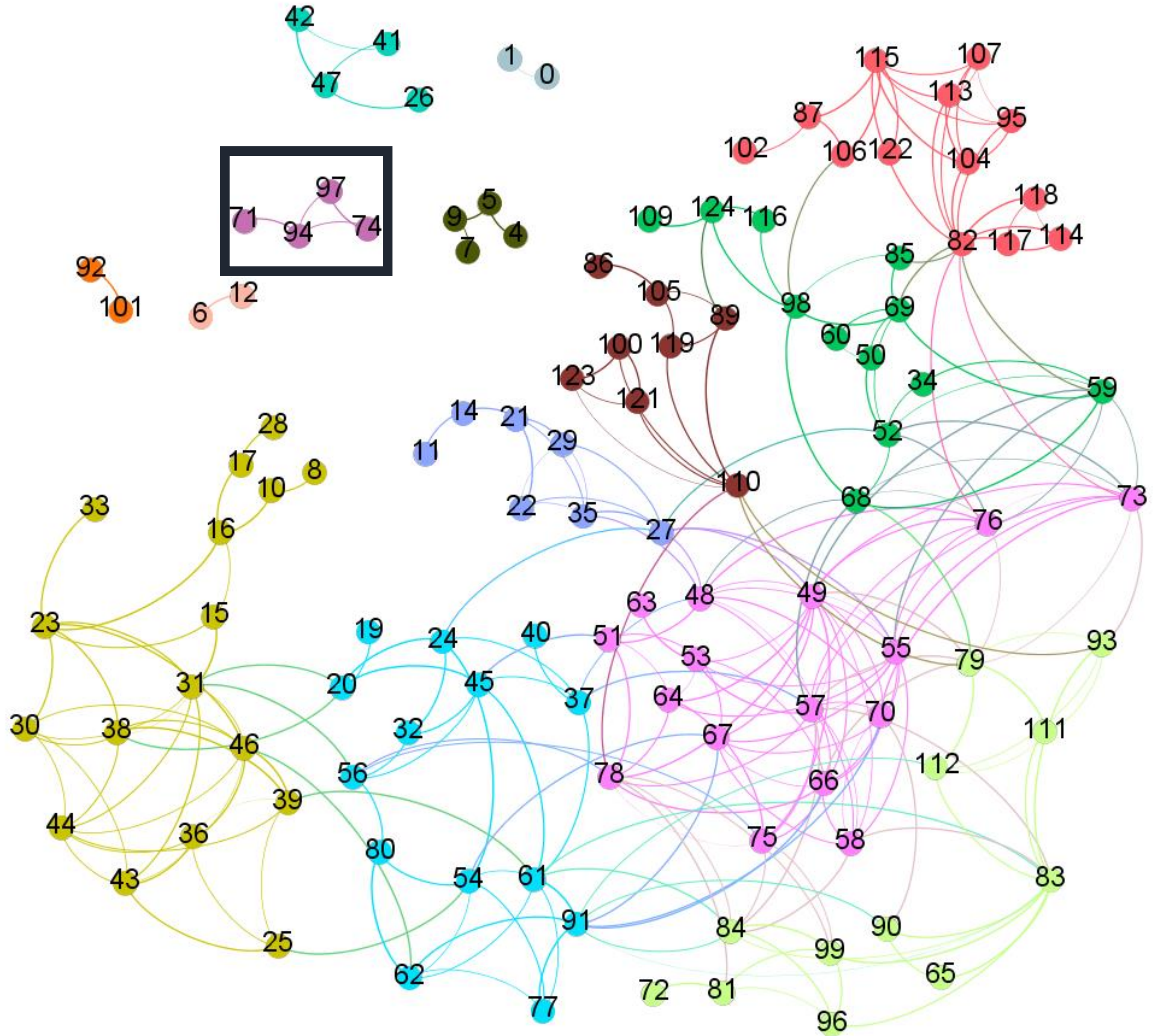
2012

Most popular thrillers

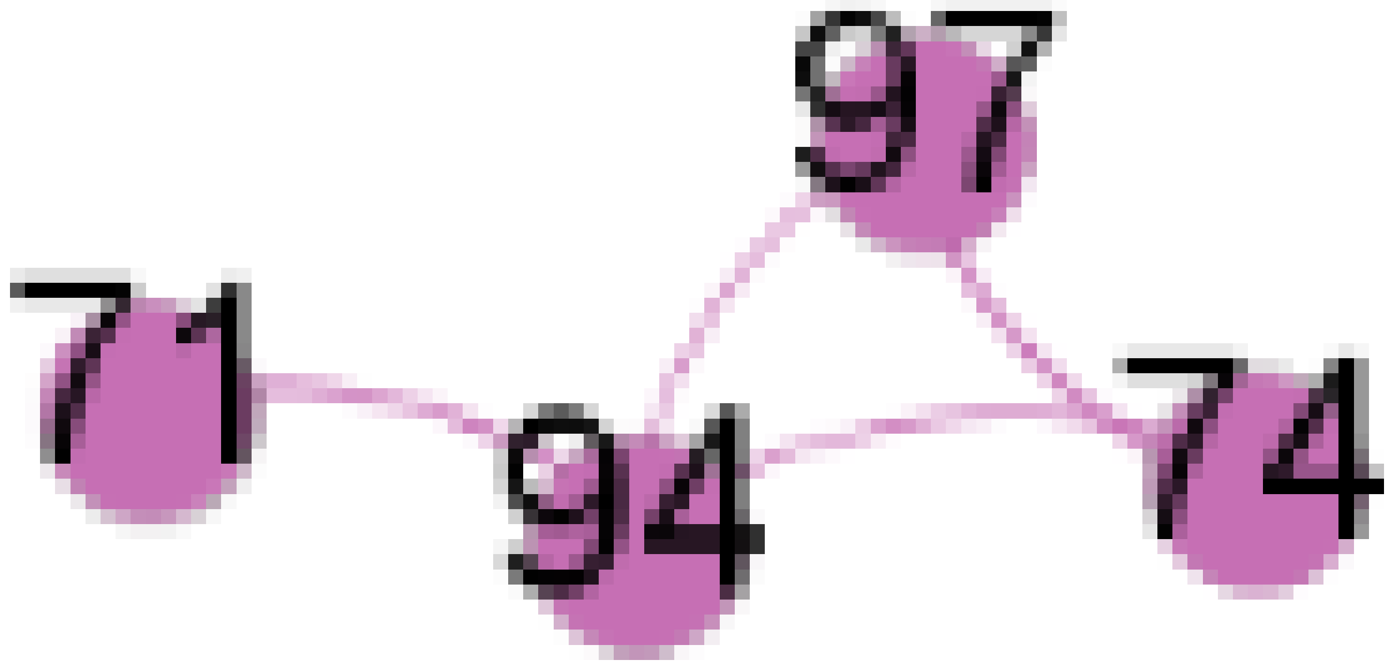
What We Already Know

NUS National University of Singapore
School of Computing

The community that likes science books especially likes suspense.



What We Already Know



71 Arrival

94 Avengers: Infinity

74 Train to Busan

97 Fast & Furious 7

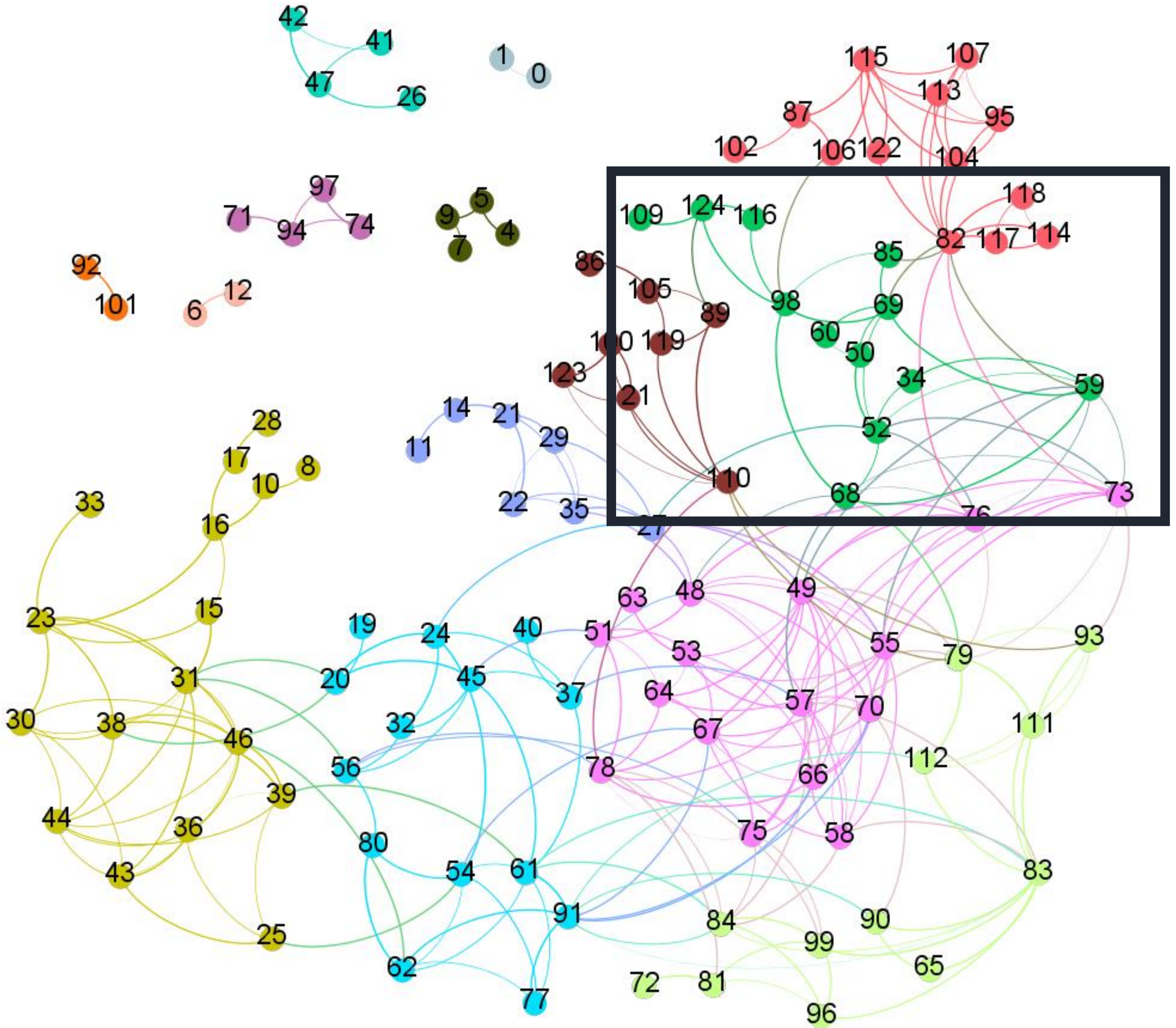
All classified because of C4, who loves science

But this group is strict, they just watch the movies in the same types

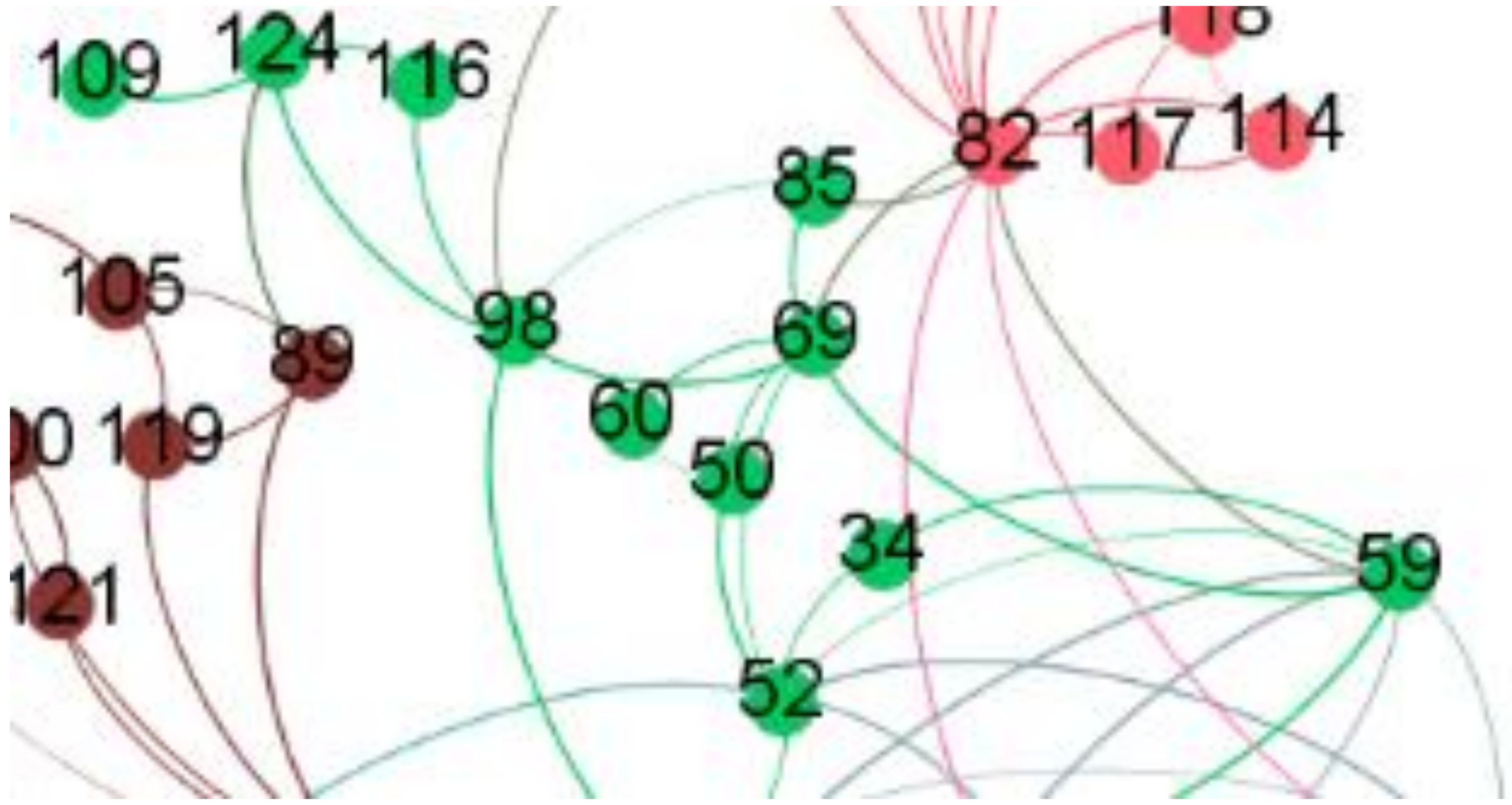
What We Do not Know

NUS National University of Singapore
School of Computing

We thought those who read about pop literature may like a variety of movies, but it turns out the different.



What We Do not Know



Black Swan （黑天鹅）

Léon（这个杀手不太冷）

You Are the Apple of My Eye （那些年,我们追过的女孩）

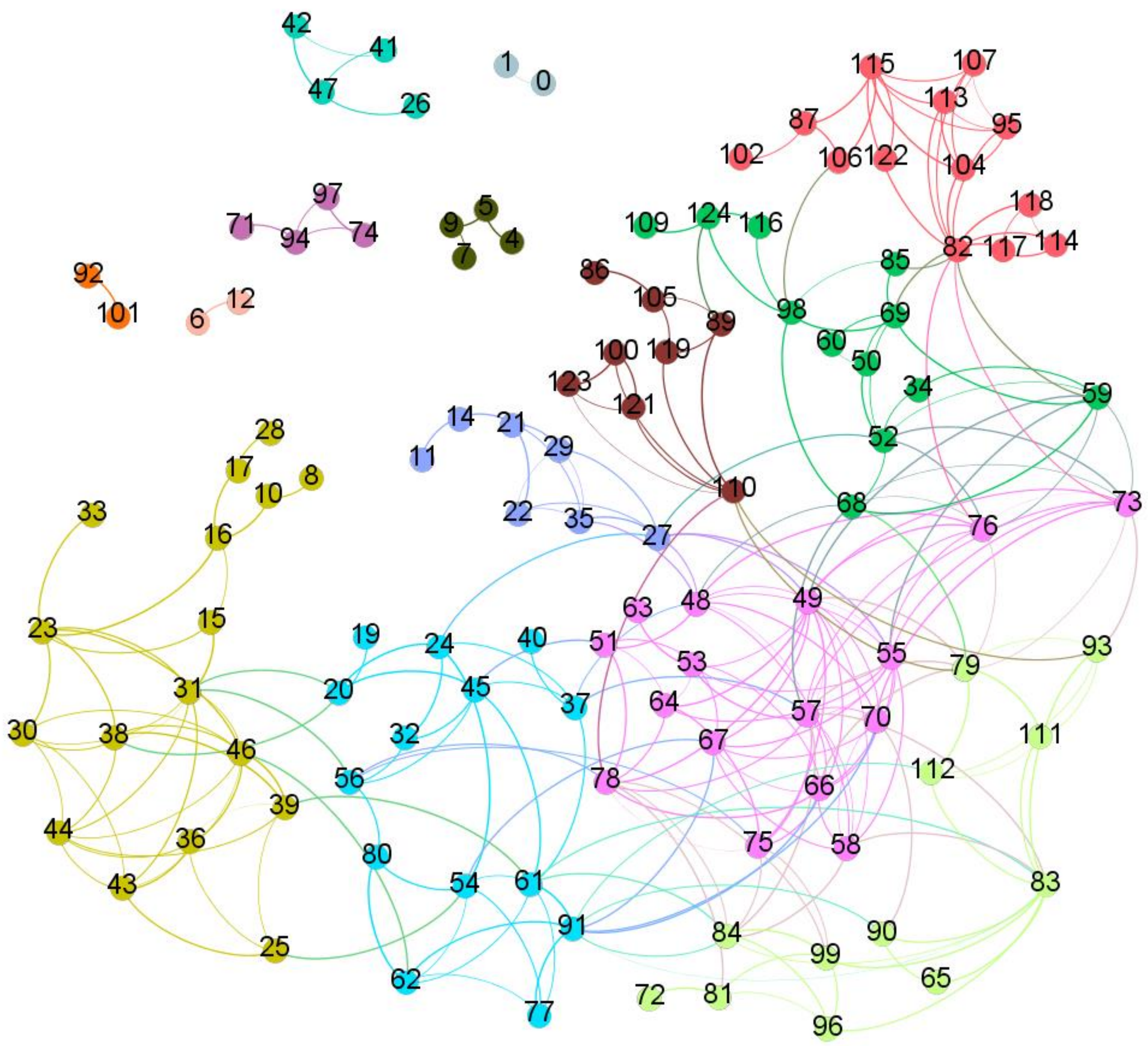
……

They don't like stimulating movies, but they like movies with a tight narrative, such as *Black Swan, Léon, You Are the Apple of My Eye.*

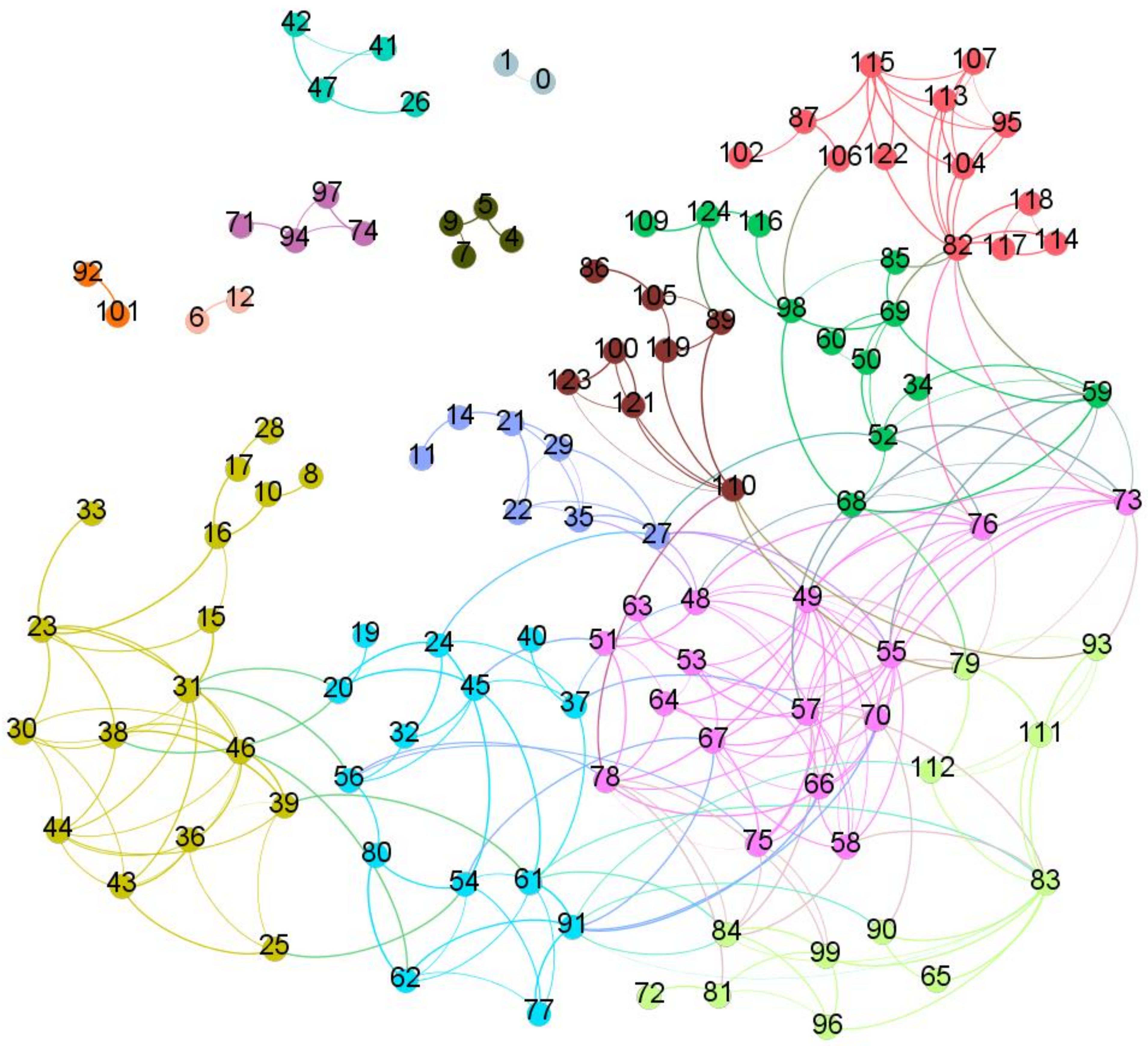
Actually, pop literature are mainly entertaining novels, such as novels from LouisCha(金庸) and Qiongyao, so a stimulating plot（刺激的情节）does not appeal to the readers.

What We Do not Know

We thought the literary community may like only a limited number of movies. In fact, they are fond of the plot, no matter the movies' popularity, and they watch the more movies.



What We Do not Know



They appear in most communities.

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14
C6	0	1	0	1	1	1	1	1	1	1	0	1	1	0

So we can analysis what movie they don' like.

Finally, we find out that they just don't prefer sci-fi movies. Like Avengers series and Cloud Atlas (云图)

What We Do not Know

The Butterfly Effect

2012

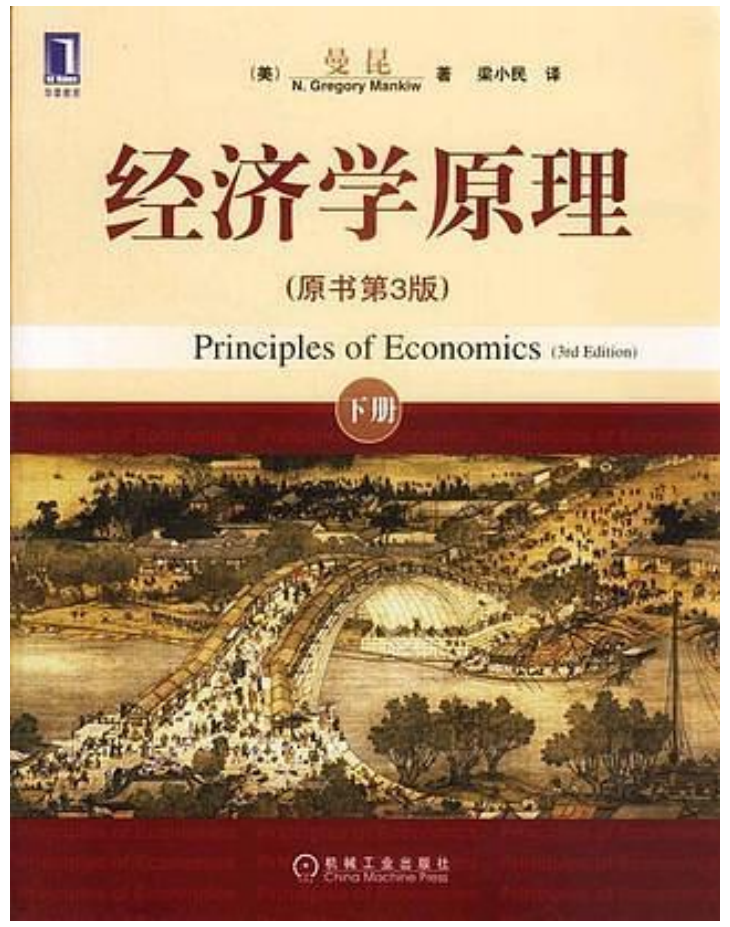
Actually those who reads about philosophy or economy like science fiction movies very much.

Even the threshold is changed ,this two node are linked together, and C1 or C5 is responsible for that.

What We Do not Know

NUS National University of Singapore
School of Computing

Reconsider the question: if a guy here reads these books, which movie he may prefer watching?

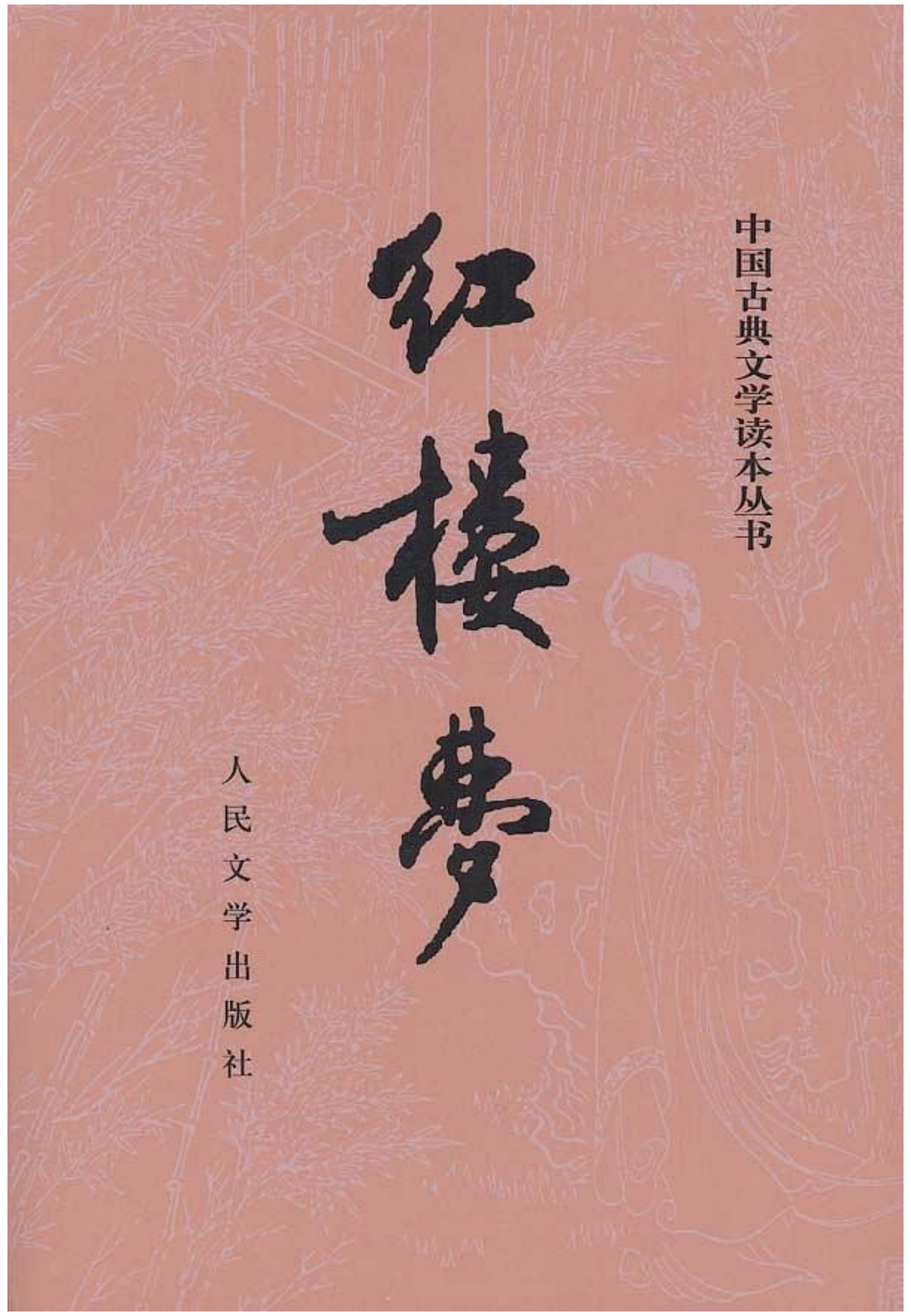


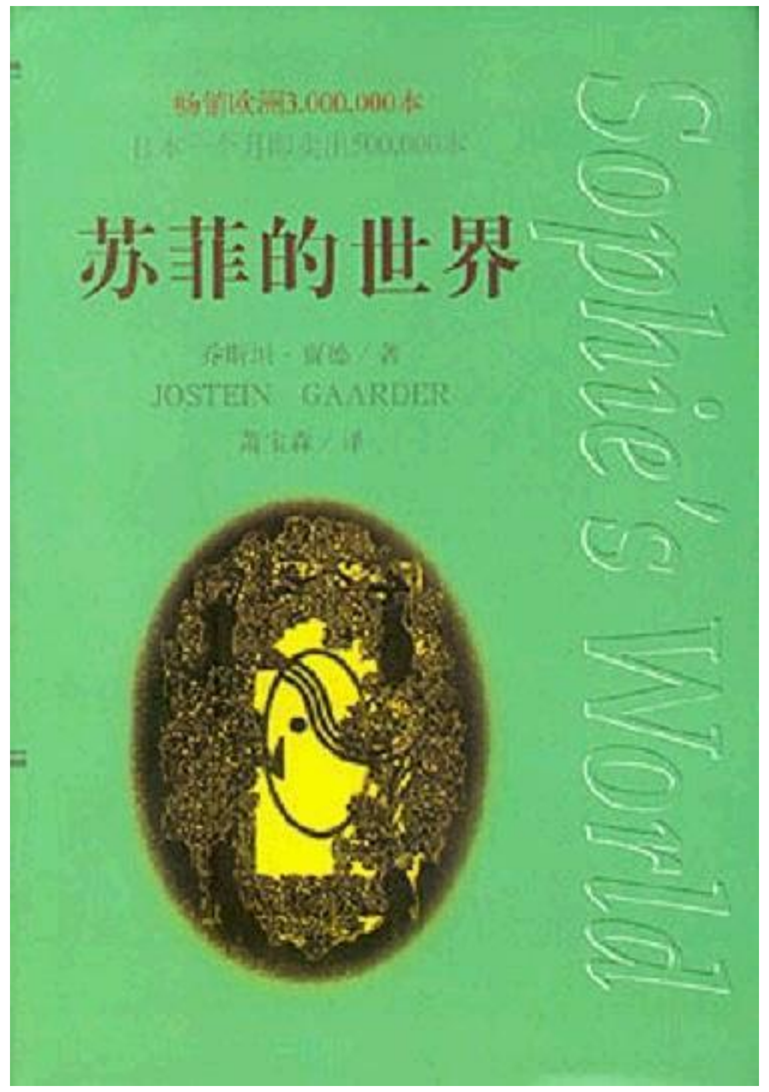
Examples

Books to Movies

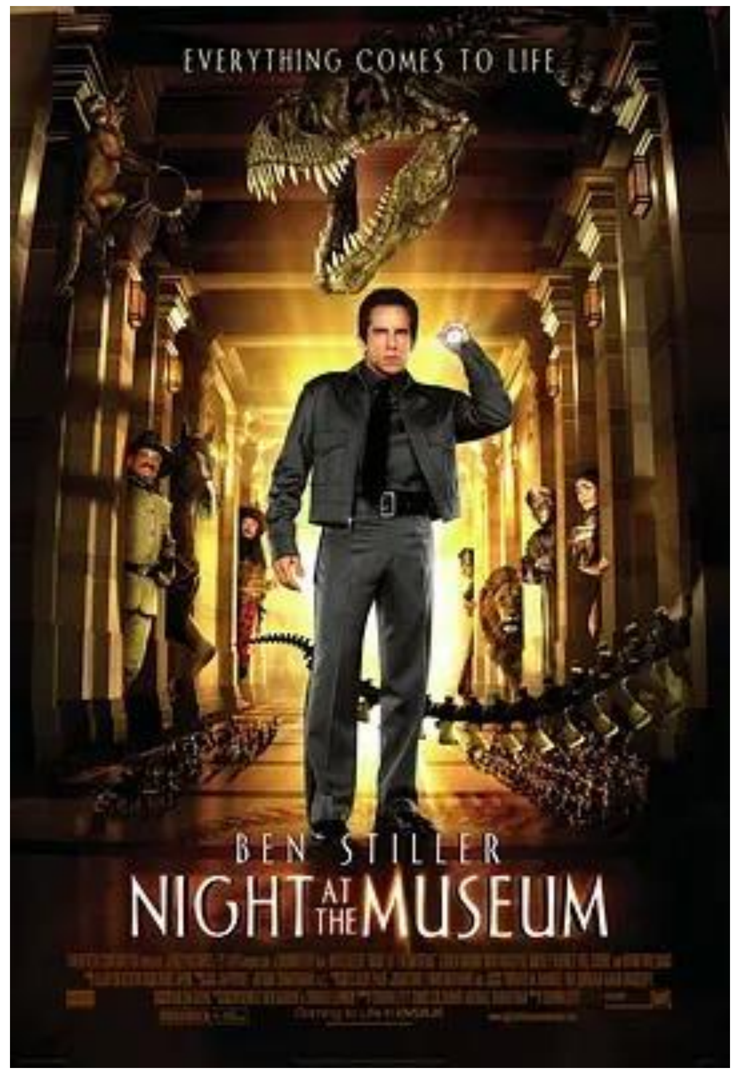
Examples

NUS
National University of Singapore
School of Computing





Examples

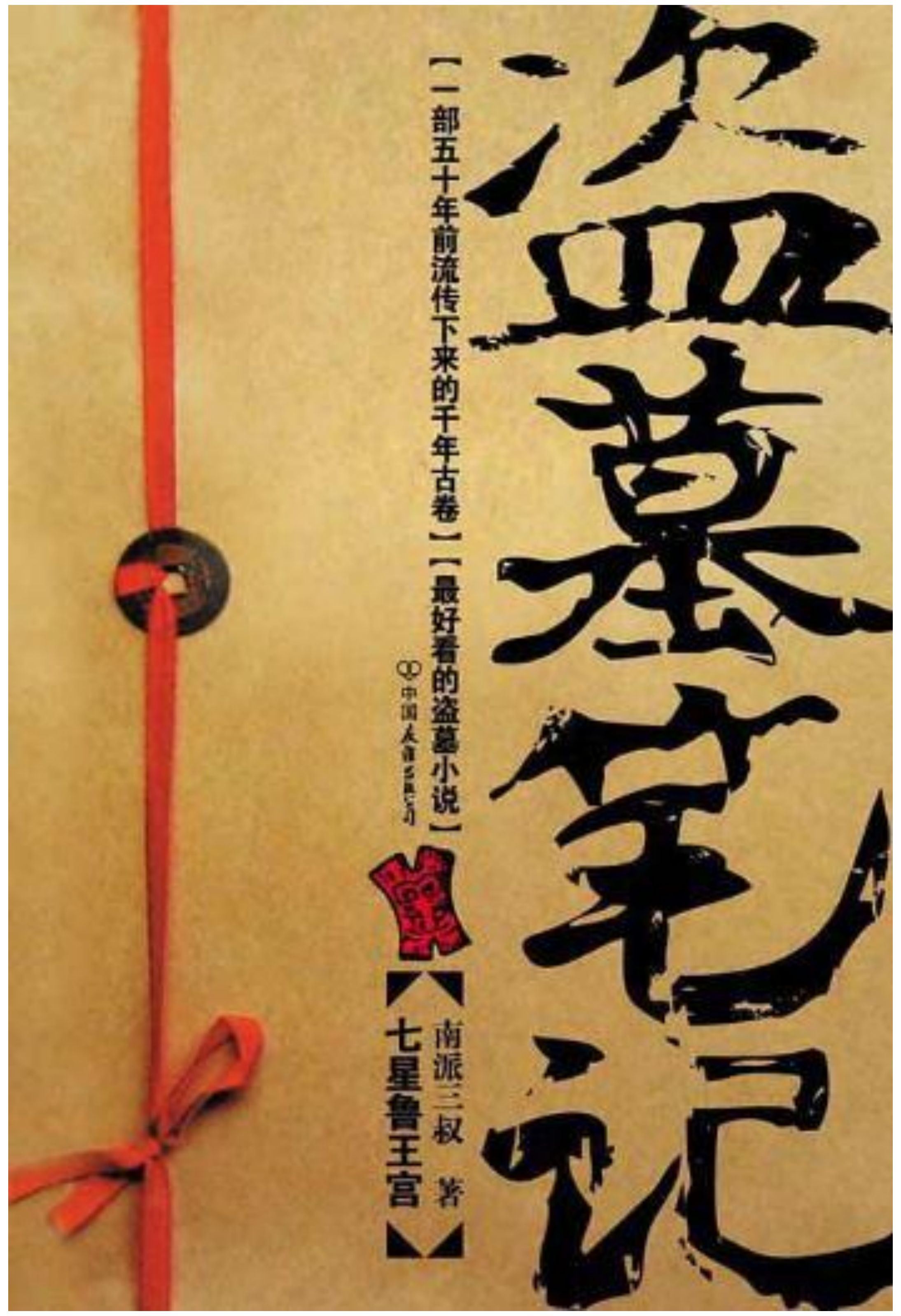

EVERYTHING COMES TO LIFE
BEN STILLER
NIGHT AT THE MUSEUM


2012·12·12
囧神贺岁
提前狂欢
人再囧途之泰囧
LOST IN THAILAND


特笑大片
西虹市首富
HELLO MR. BILLIONAIRE
闫非 彭大魔 电影作品
2018.7.27
是金子总会花光的


9.30 一起哈哈哈
夏洛特烦恼
Goodbye Mr. Loser
开心麻花

Examples


盗墓笔记
【一部五十年前流传下来的千年古卷】【最好看的盗墓小说】
南派三叔 著
【七星鲁王宫】

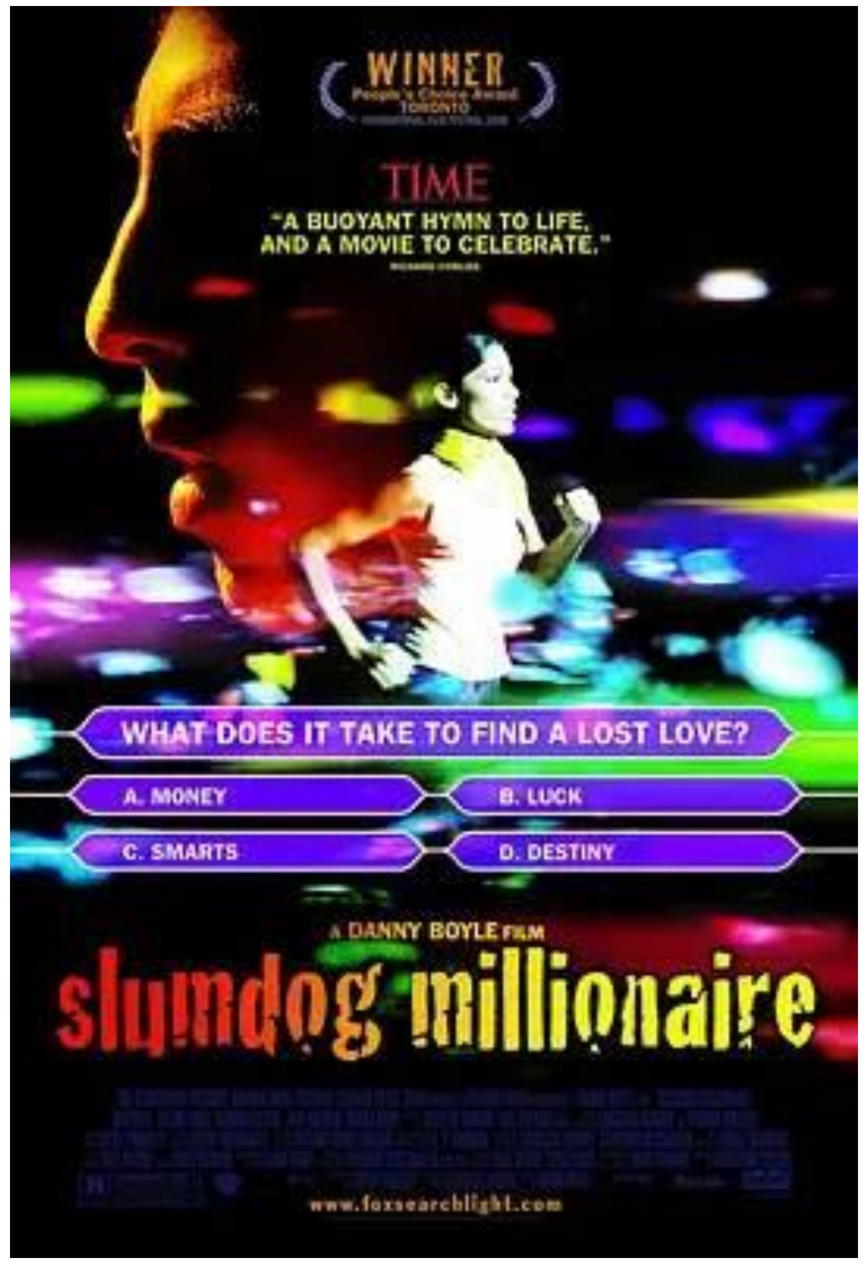

村 上 春 树
一 首 感 伤 的 青 春 恋 曲
挪威的森林
（全译本）林少华 译
上海译文出版社


白夜行
びゃくやこう
〔日〕东野圭吾 著 刘姿君 译

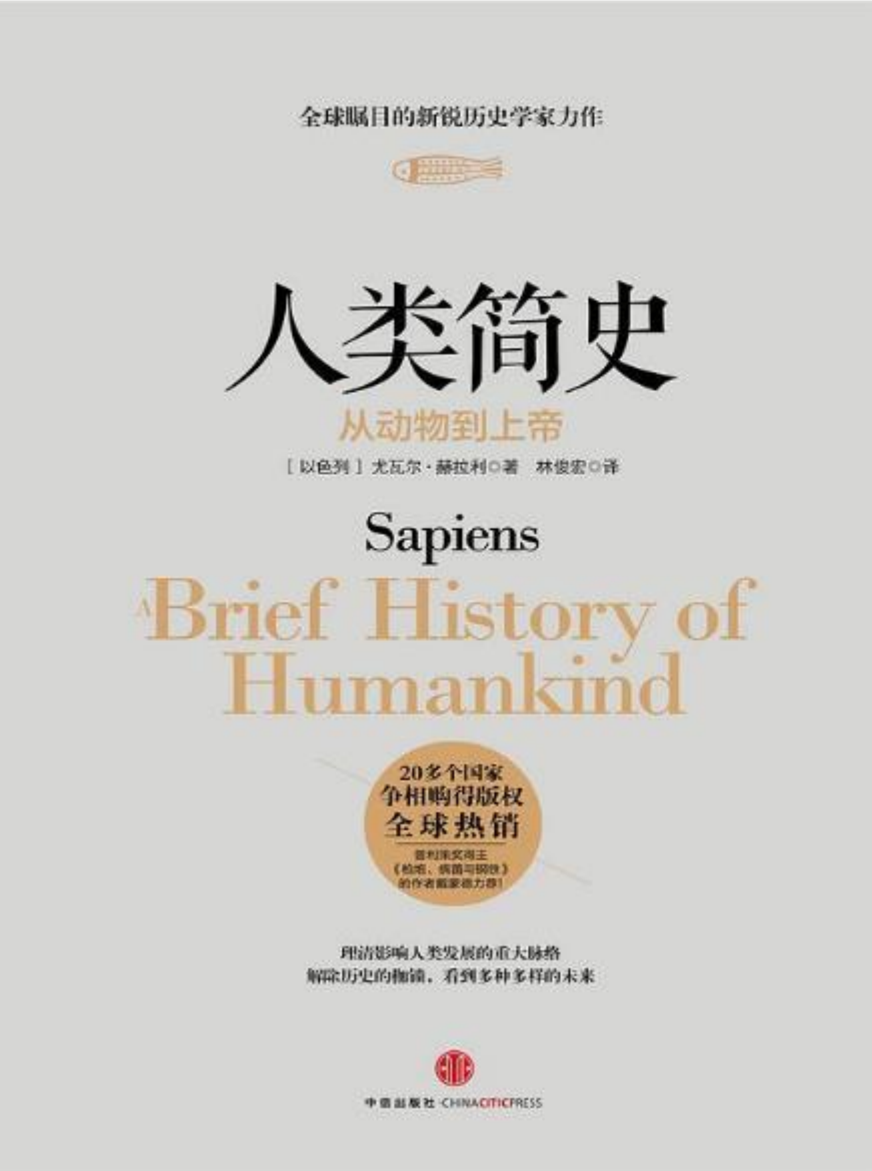
Examples

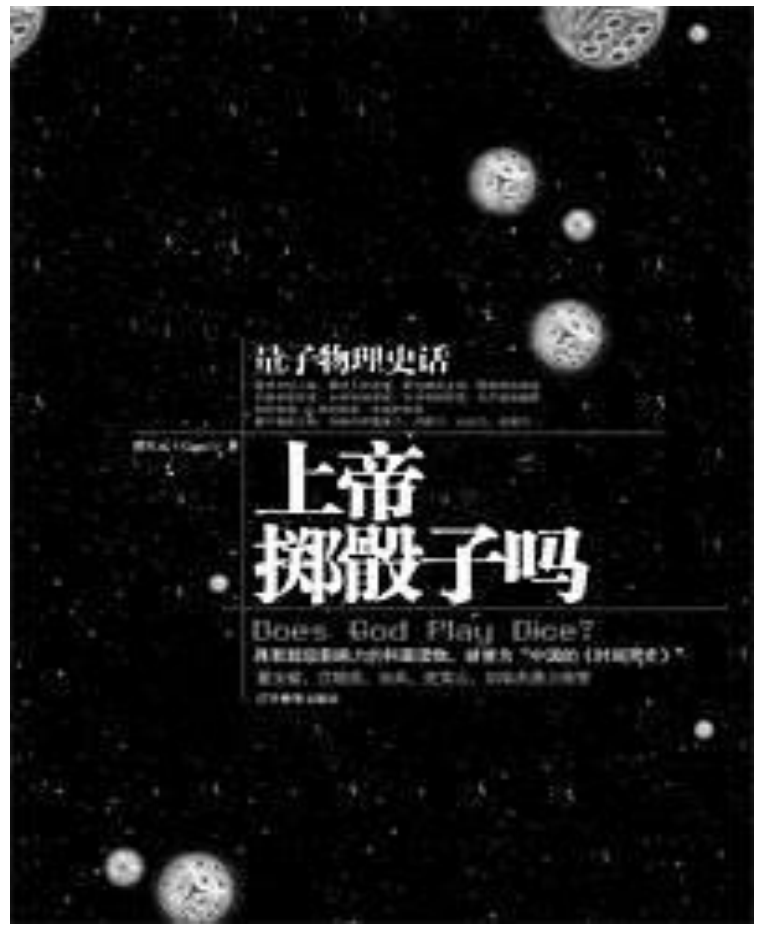

我不是药神
DYING TO SURVIVE
2018.7.6

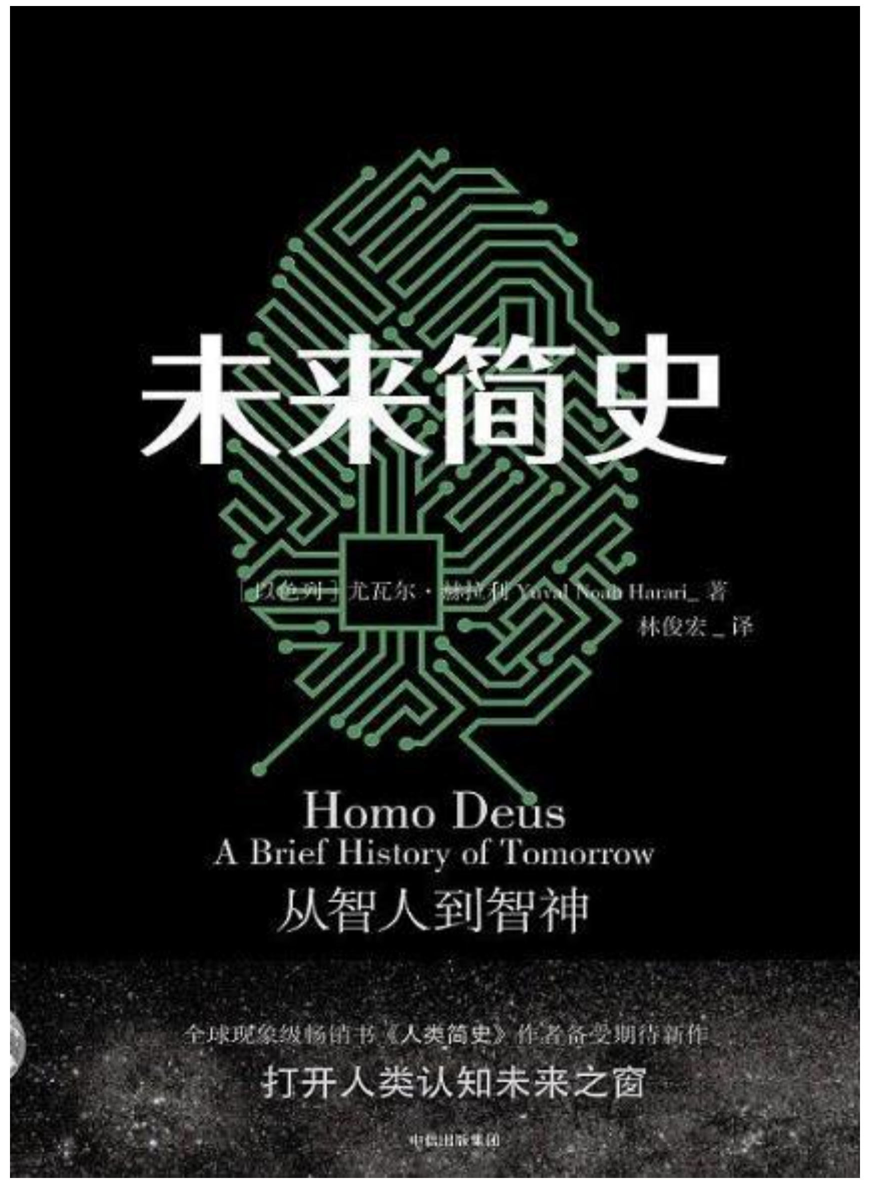

摔跤吧！爸爸
DANGAL
5月5日 先摔为敬


WINNER
TIME
"A BUOYANT HYMN TO LIFE, AND A MOVIE TO CELEBRATE."
WHAT DOES IT TAKE TO FIND A LOST LOVE?
A. MONEY
B. LUCK
C. SMARTS
D. DESTINY
A DANNY BOYLE FILM
slumdog millionaire
www.foxsearchlight.com

Examples







Particular about movies

Examples

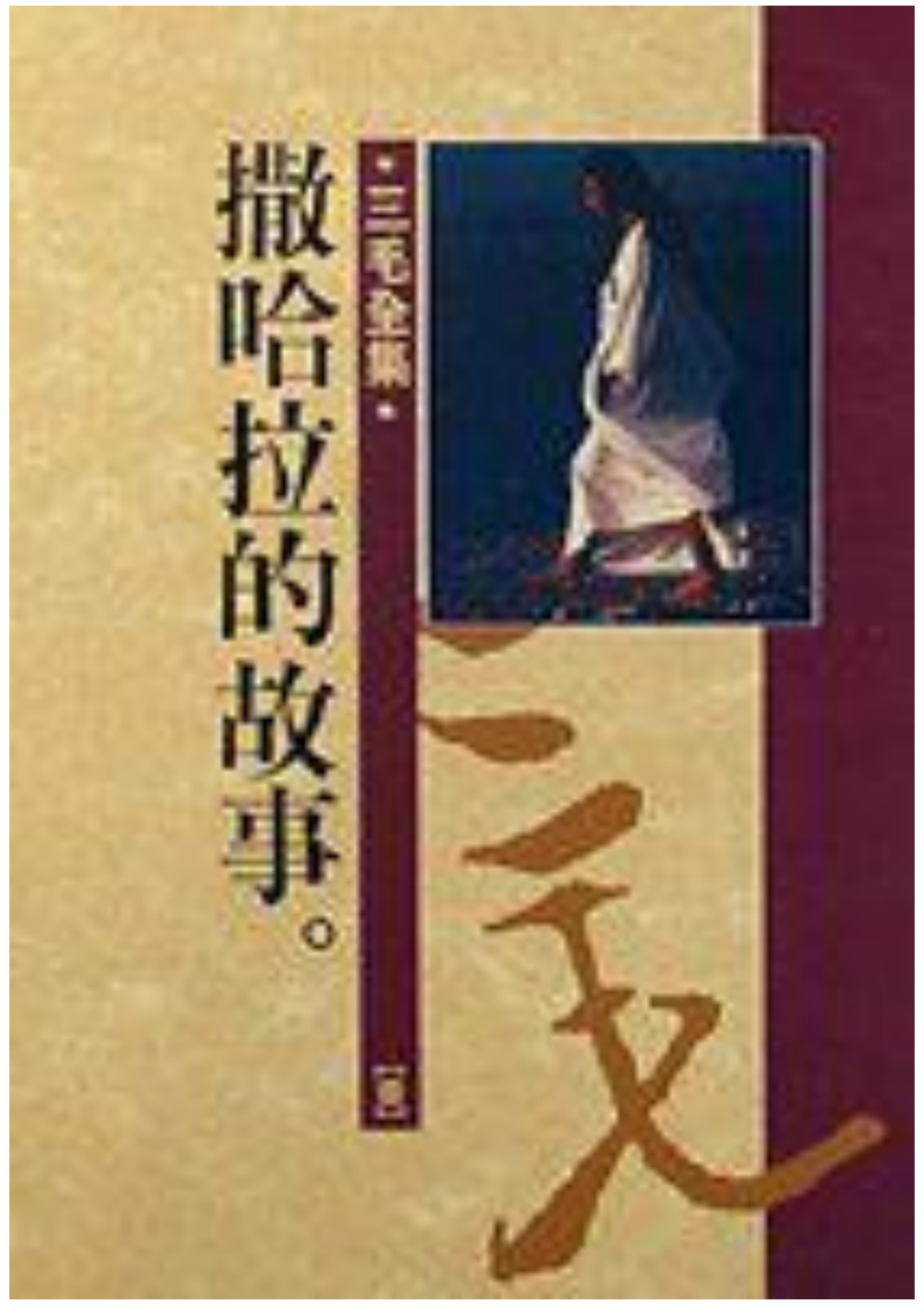

马修 麦康纳
安妮 海瑟薇
杰西卡 查斯坦
迈克尔 凯恩
星际穿越

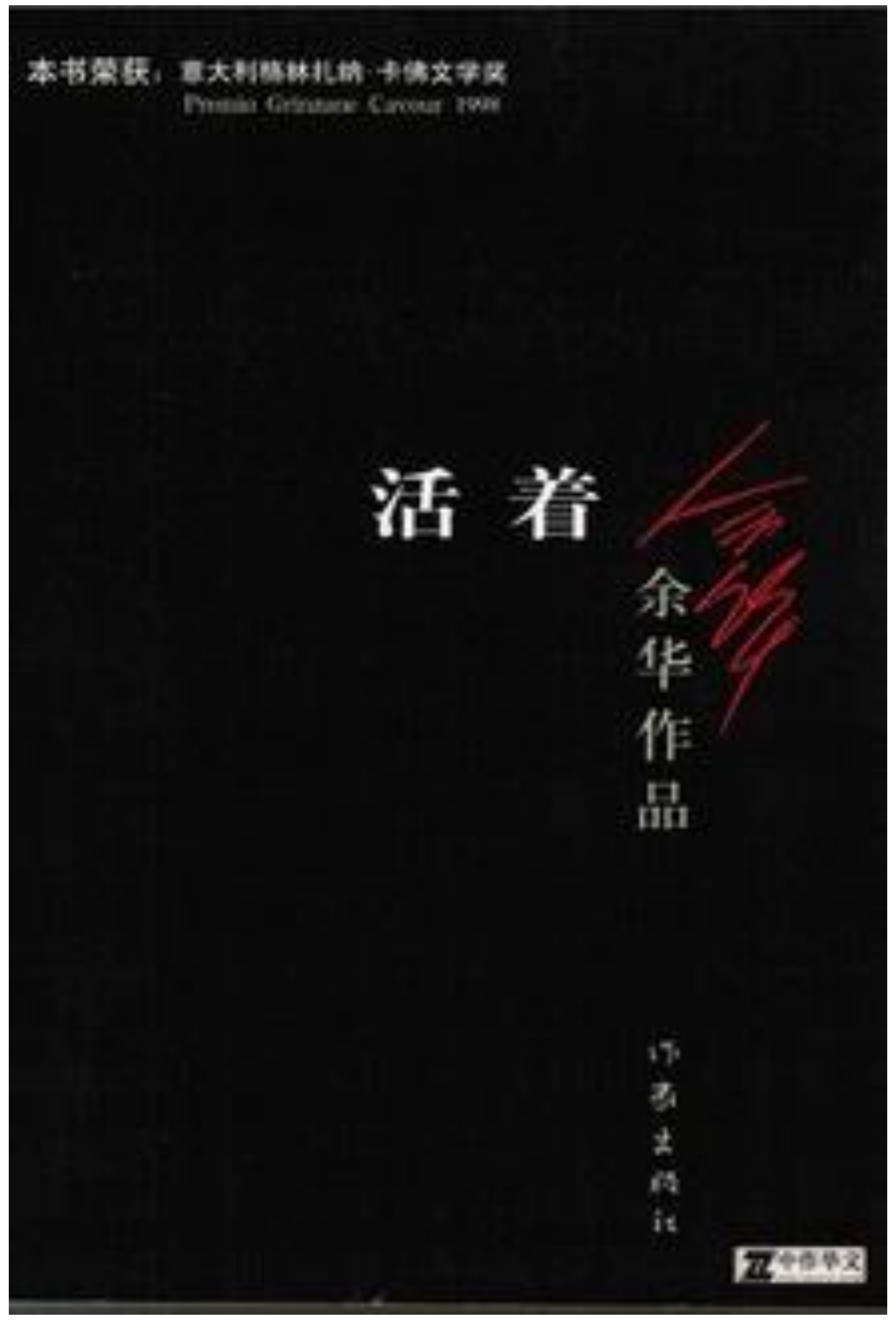

THE WANDERING EARTH
流浪地球
2019·大年初一
IMAX 3D

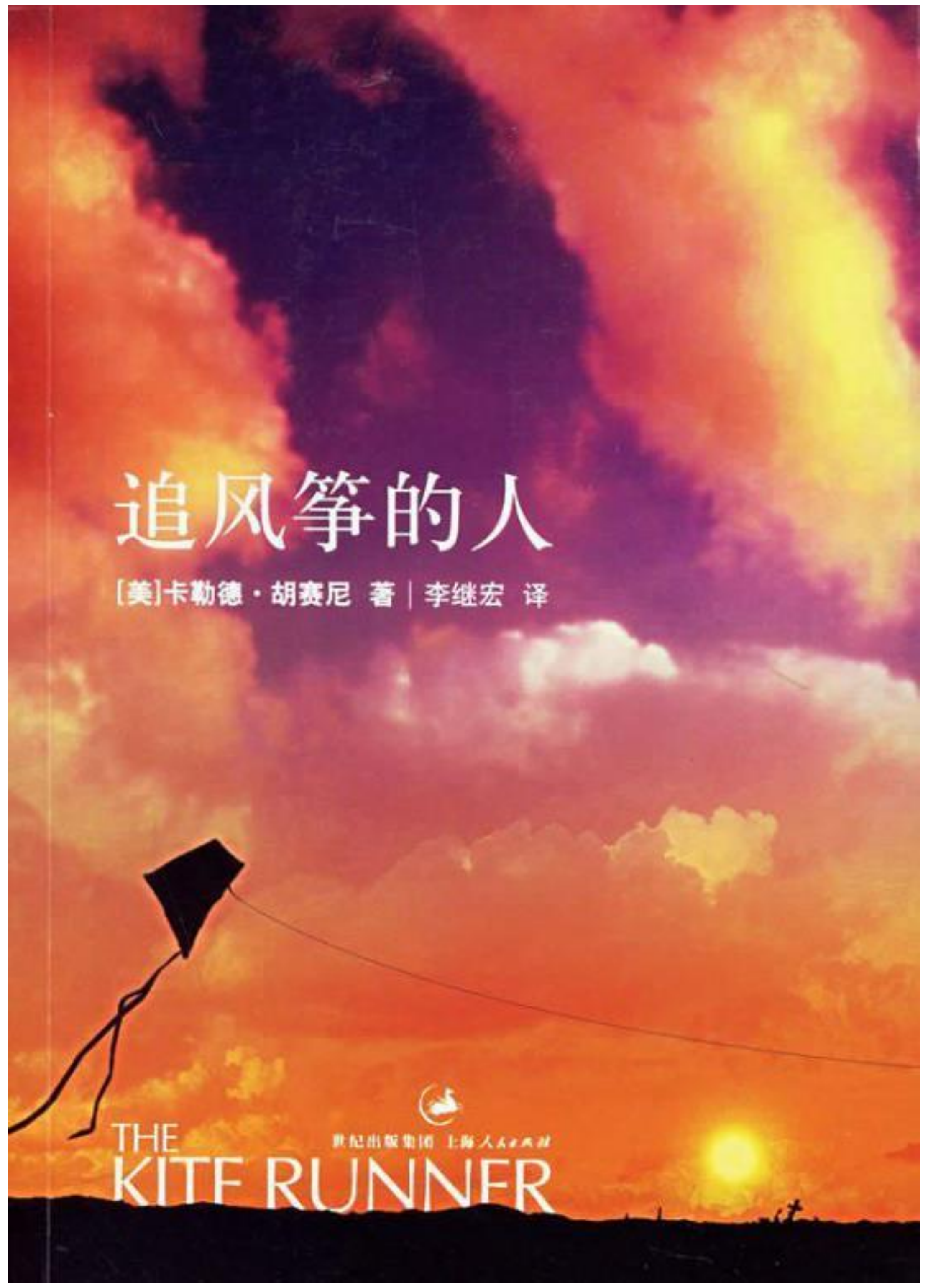

为生者权
06/14
为死者言
秦明
生死语者


十年布局
巅峰一役
MARVEL STUDIOS
复仇者联盟3
5月11日 空前集结
3D / IMAX 3D

Examples


撒哈拉的故事。
·三毛全集·

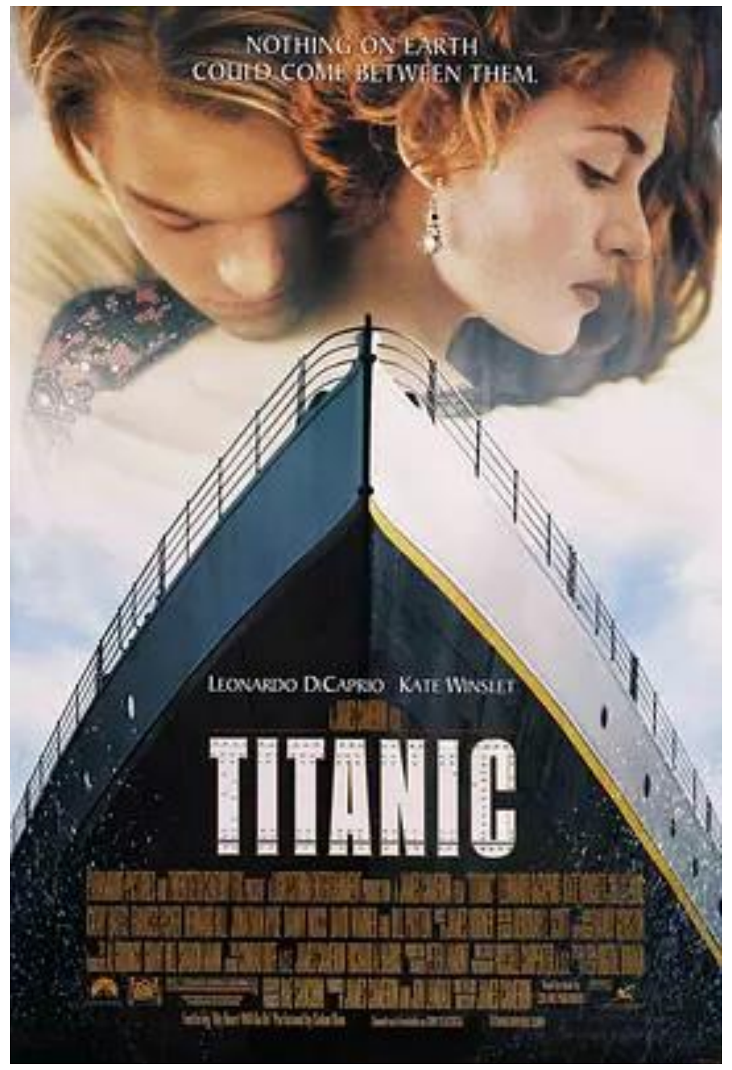

本书荣获：意大利格林扎纳·卡佛文学奖
活 着
余华作品


追风筝的人
[美]卡勒德·胡赛尼 著 | 李继宏 译
THE KITE RUNNER

Examples

Examples


FROM THE DIRECTOR OF THE BUCKET LIST
STAND BY ME AND WHEN HARRY MET SALLY...
YOU NEVER FORGET YOUR FIRST LOVE.
FLIPPED
AUGUST


NOTHING ON EARTH
COULD COME BETWEEN THEM.
LEONARDO DICAPRIO KATE WINSLET
TITANIC


千与千寻
SPIRITED AWAY
06.21


FEAR CAN HOLD YOU PRISONER.
HOPE CAN SET YOU FREE.
TIM ROBBINS MORGAN FREEMAN
THE SHAWSHANK REDEMPTION

NUS National University of Singapore
School of Computing

Q&A

Website

Thank you

Books to Movies