Natural Language Understanding, Generation, and Machine Translation

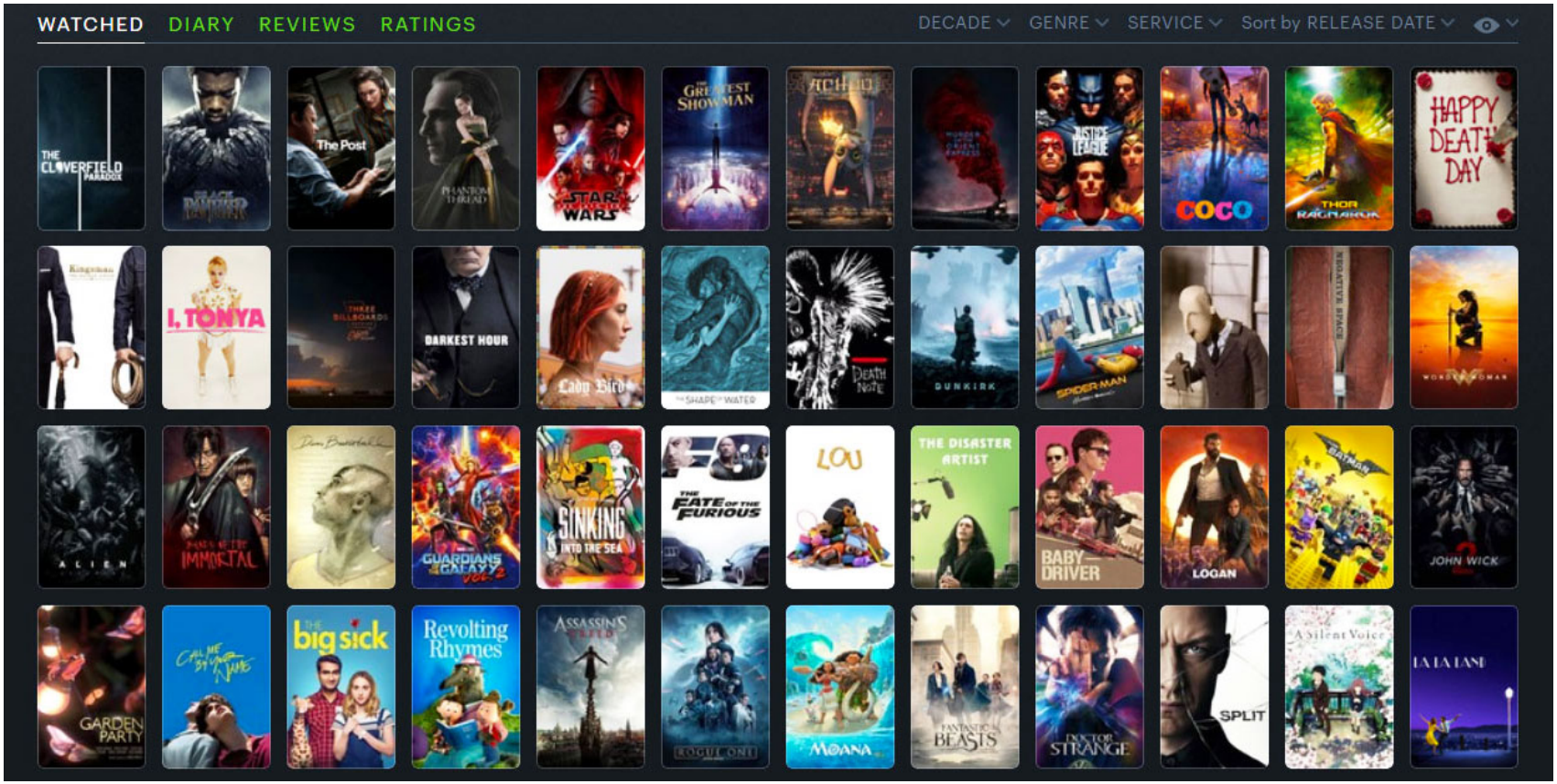
Lecture 25: Movie Summarization

Frank Keller
School of Informatics
University of Edinburgh

Joint work with Pinelopi Papalampidi, Lea Frermann, and Mirella Lapata.

Who doesn't Like Movies?

Movies are Everywhere!



- Netflix has **6,000–7,000** movies, series, and shows
- Amazon has **24,000** movies and **2,100** shows to choose from
- The average person watches TV **≈3.5h** per day

Viewers Need Help!

- The average[1] household has three streaming services
- 70% agree they often struggle to figure out what to watch
- Viewers need more than 15 min to decide what to watch

[1] Statistics from Likewise Inc.

Viewers Need Help!

- Give me an **alternative trailer** or a **10 minute movie summary.**
- Show me **action-packed** shots or **scenes with female leads.**
- I want to see all **funny/happy** shots.

Movie Analysis

Previous approaches:

- **Text-only:** character-based networks, **screenplay** summaries

  (Gorinski and Lapata, 2015, Papalampidi et al., 2020)

- **Video-only:** focus on **isolated** video clips

  (Tapaswi et al., 2016; Rohrbach et al., 2015, Lei et al., 2019)

Movie Analysis

Previous approaches:

- **Text-only:** character-based networks, **screenplay** summaries

  (Gorinski and Lapata, 2015, Papalampidi et al., 2020)

- **Video-only:** focus on **isolated** video clips

  (Tapaswi et al., 2016; Rohrbach et al., 2015, Lei et al., 2019)

Our approach:

- **Multi-modal** input: screenplay + movie video
- Summaries derived from **full-length movies**
- Based on **Turning Point (TP)** identification

Movie Summarization

Challenges

- Movies are **long** (≈2 hours), **complex** stories
- Position biases or importance and diversity heuristics **do not lead to meaningful summaries**
- There are **gold-standard summaries** available, huge effort to collect video annotations

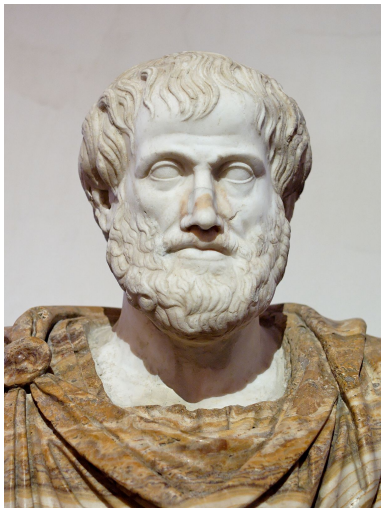
Movie Summarization

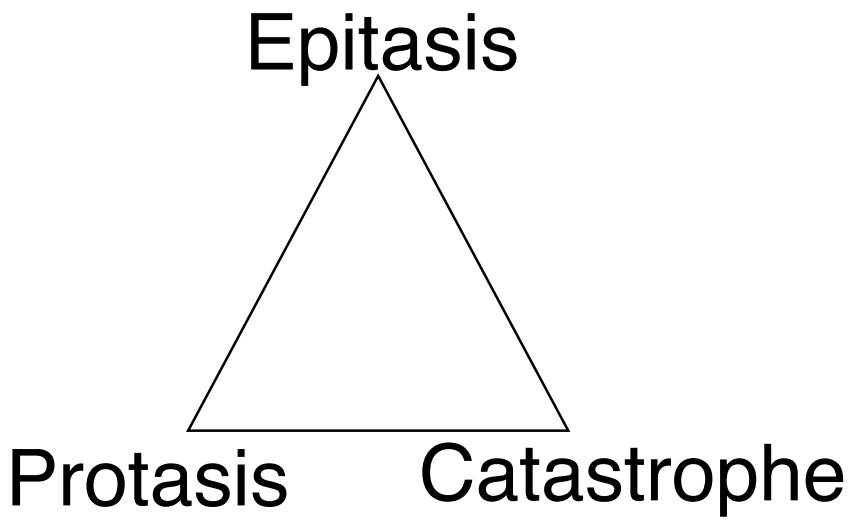
Challenges

- Movies are **long** (≈2 hours), **complex** stories
- Position biases or importance and diversity heuristics **do not lead to meaningful summaries**
- There are **gold-standard summaries** available, huge effort to collect video annotations

Our **hypothesis:** a good movie summary consists of scenes that are **turning points** in the story.

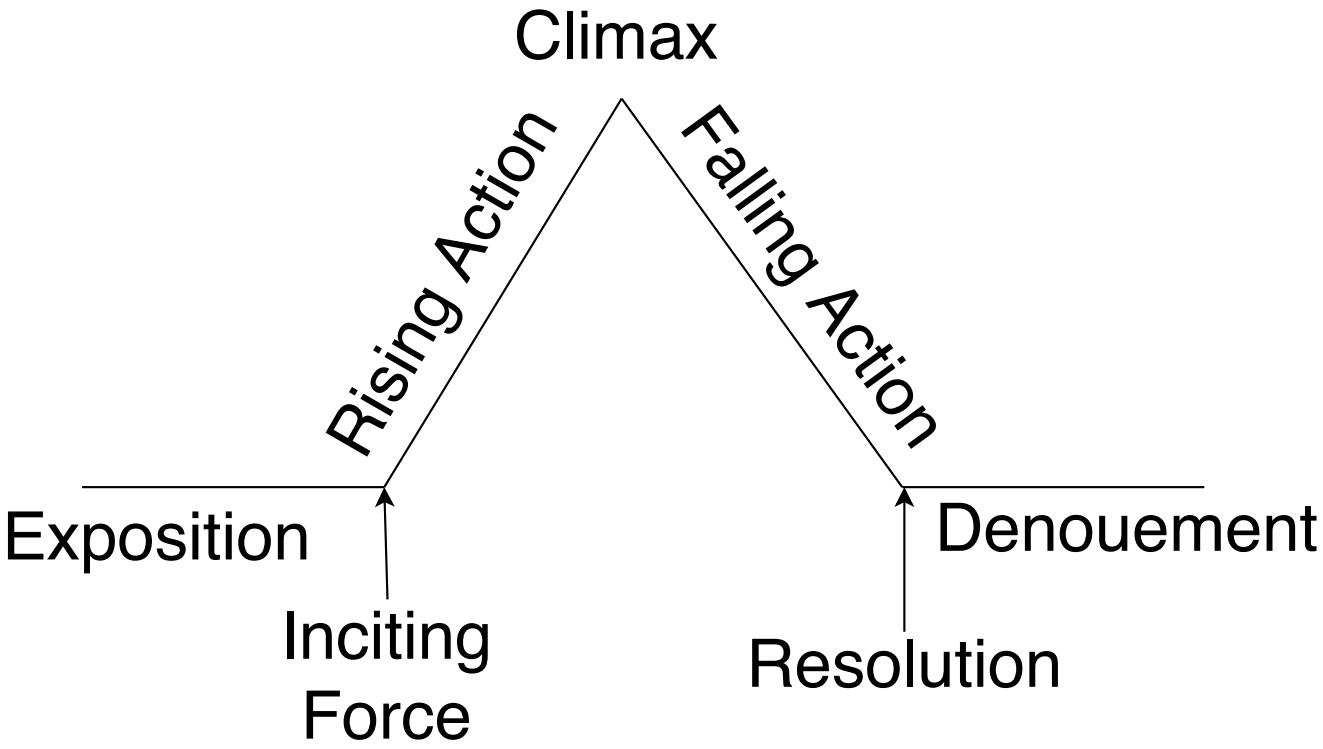
What Are Turning Points?





Aristotle's triangle (Pavis, 1998)

What Are Turning Points?



Freytag's (1896) pyramid

What Are Turning Points?

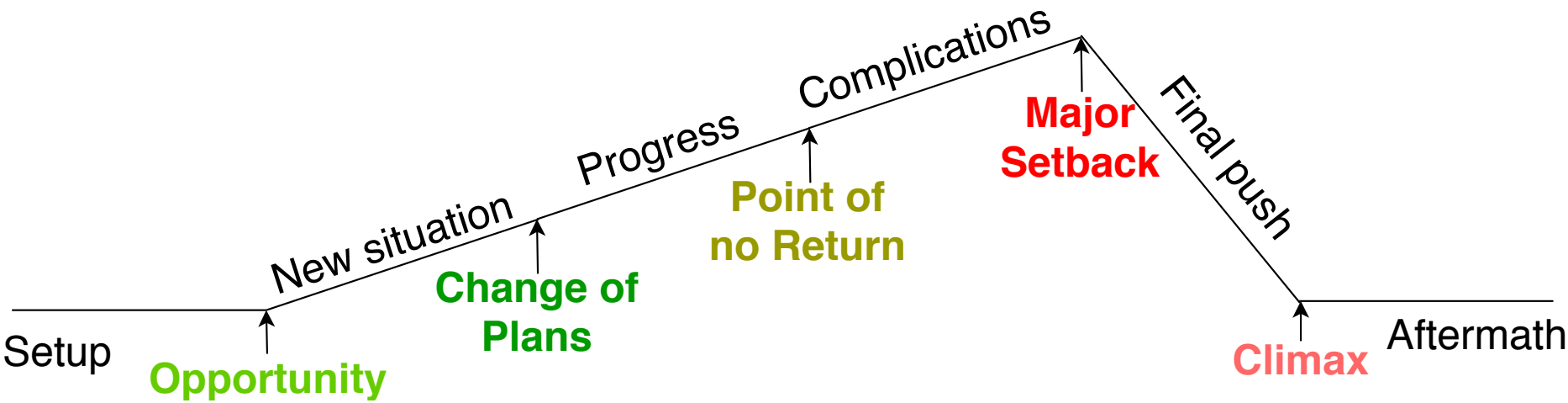
Turning points (TPs) are important narrative moments (Thompson, 1999):

- determine the **progression** of the plot
- segment the narrative into **thematic units**

What Are Turning Points?

Turning points (TPs) are important narrative moments (Thompson, 1999):

- determine the **progression** of the plot
- segment the narrative into **thematic units**
- In this work we adopt the **modern scheme** of Hague (2017)



Let's Look at an Example!

The Shining is a 1980 psychological horror film produced and directed by Stanley Kubrick is based on Stephen King's 1977 novel.

The Opportunity

Introductory event that occurs after the presentation of the setting and the background of the main characters.

Change of Plans

Event where the main goal of the story is defined. From this point on, the action begins to increase.

Point of No Return

Event that pushes the main character(s) to fully commit to their goal.

Major Setback

Event where everything falls apart (temporarily or permanently).

Climax

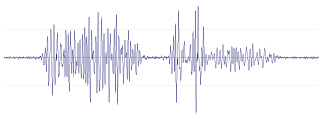
Final event of main story, moment of resolution and “biggest spoiler”.

Summarization: Problem Formulation

Screenplay

Video

Audio

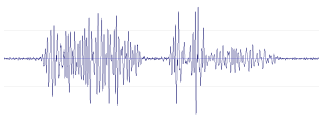


Summarization: Problem Formulation

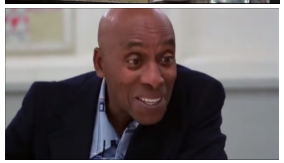
Screenplay

Video

Audio



TP1



TP2

TP3

TP4

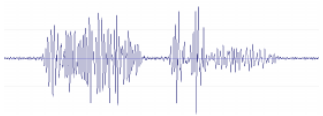
TP5

Summarization: Problem Formulation

Screenplay

Video

Audio



TP1

TP2

TP3

TP4

TP5

Video Summary

Graph-based Model for TP Identification

We model movies as **sparse graphs,** where:

- Nodes $\rightarrow$ scenes
- Edges $\rightarrow$ similarity between scenes

Why do we use graphs?

- Better **contextualization:** represent **complex interactions** between events (e.g., substories, flashbacks)
- Better **navigation:** yields **interpretable** sparse connections

Graph-based Model for TP Identification

scene representations

s_1

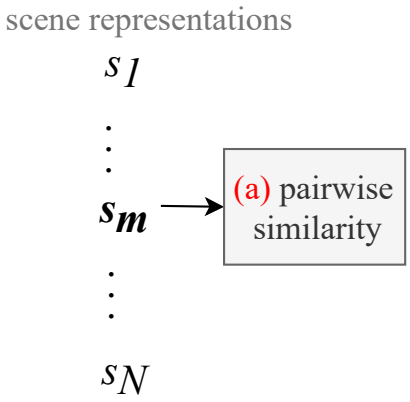
$\vdots$

$\boldsymbol{s_m}$

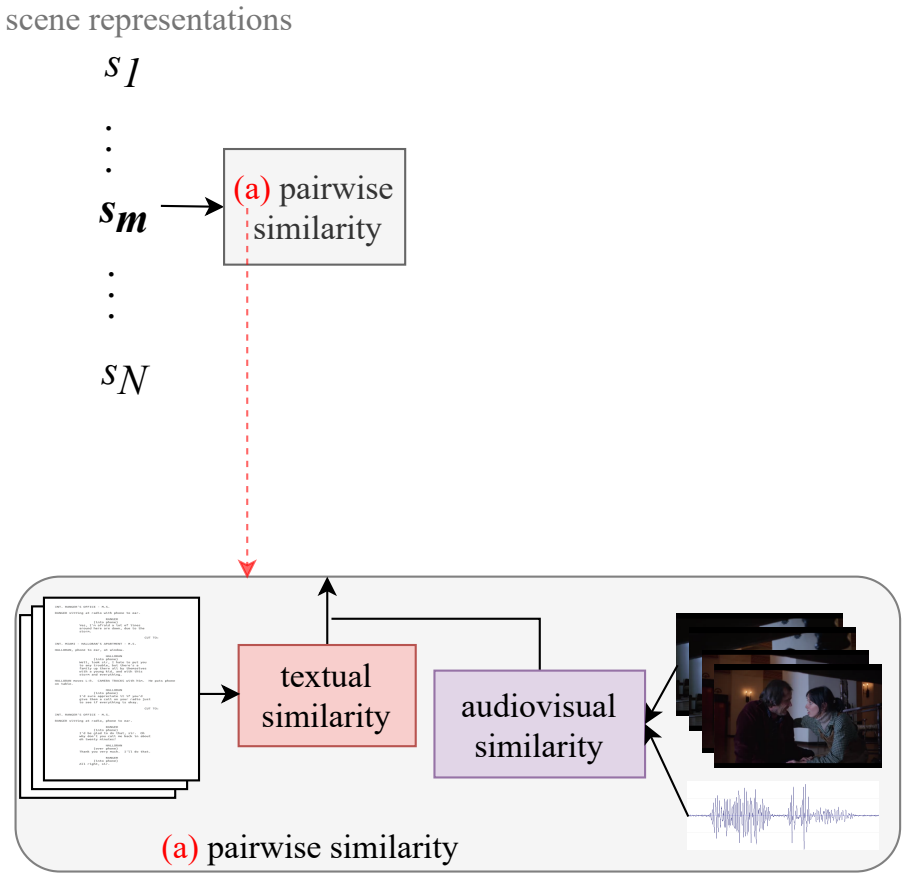
$\vdots$

s_N

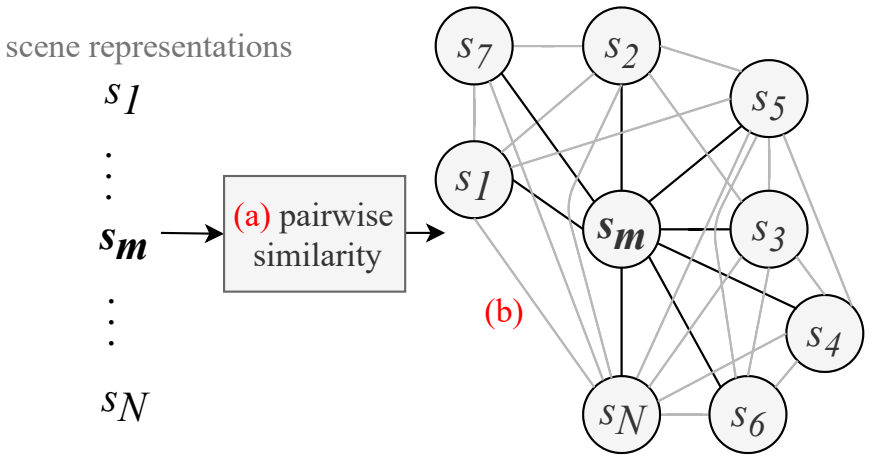
Graph-based Model for TP Identification



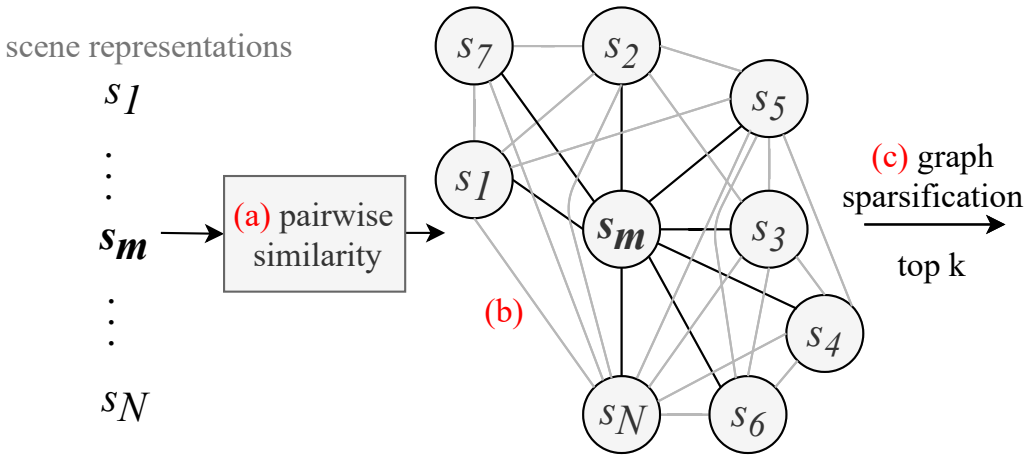
Graph-based Model for TP Identification



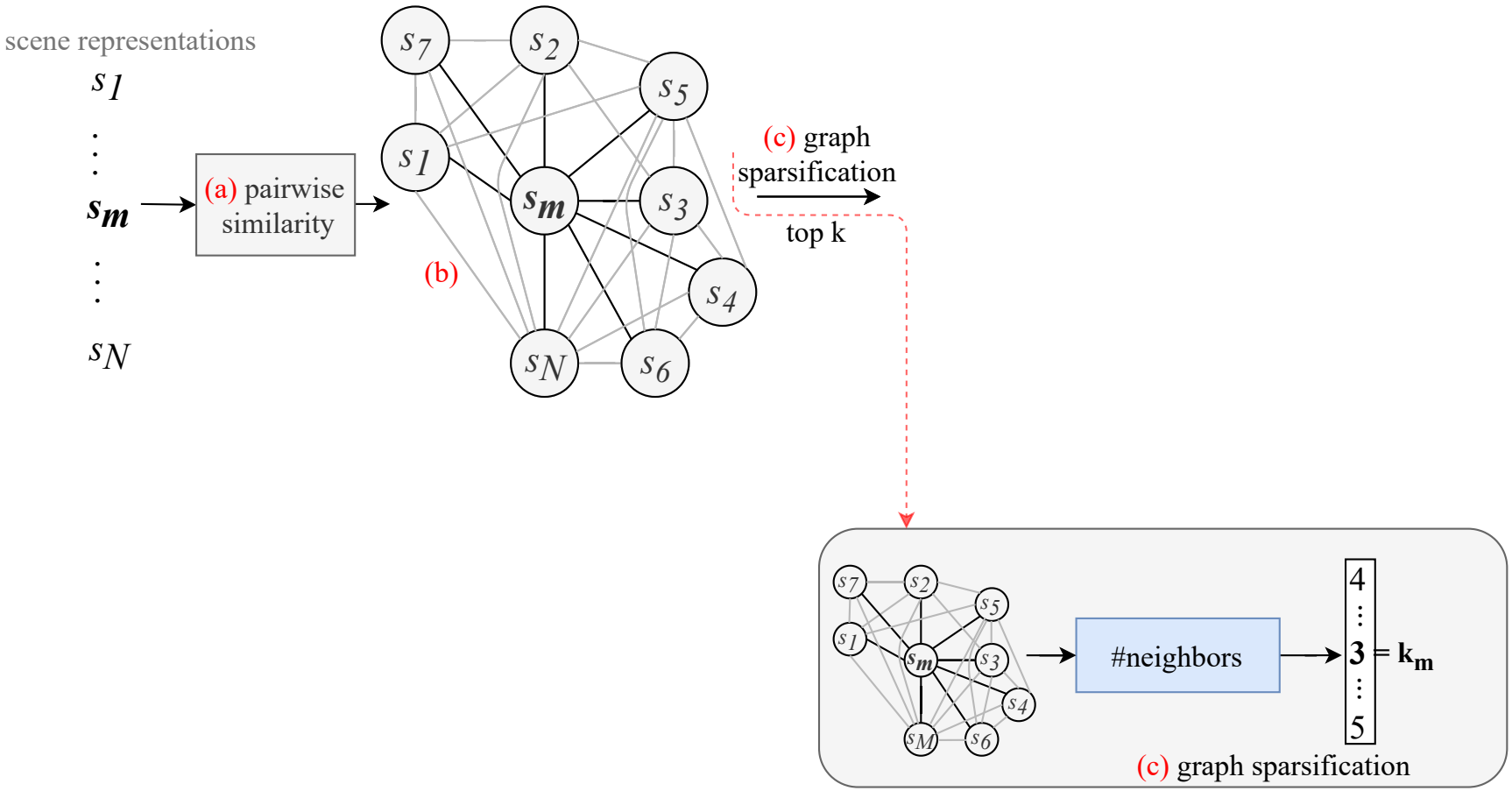
Graph-based Model for TP Identification



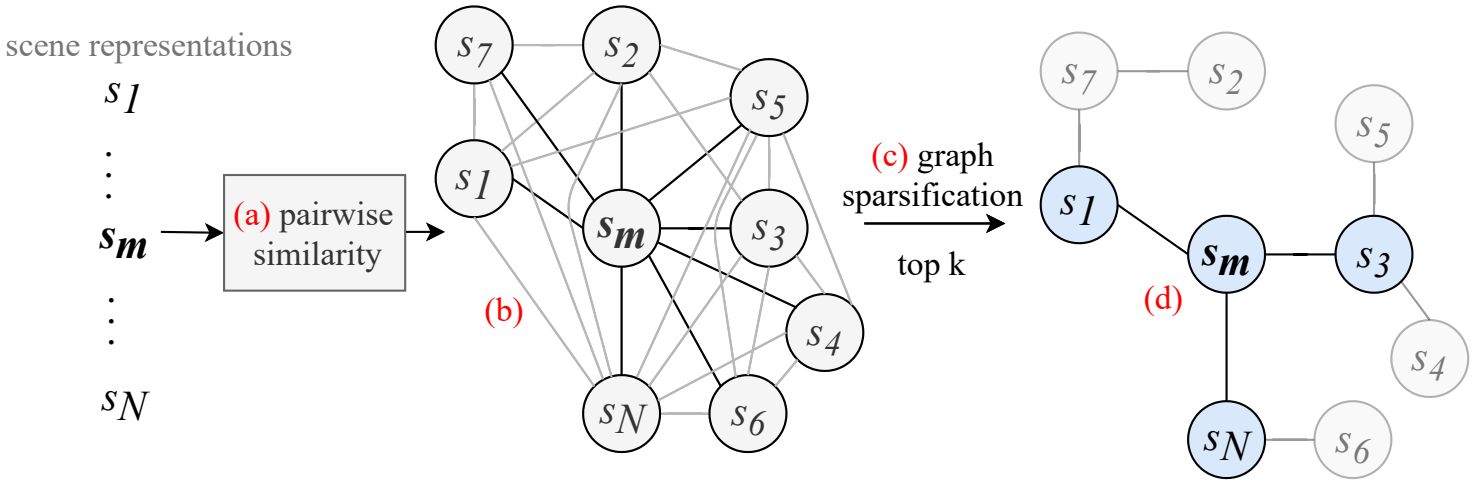
Graph-based Model for TP Identification



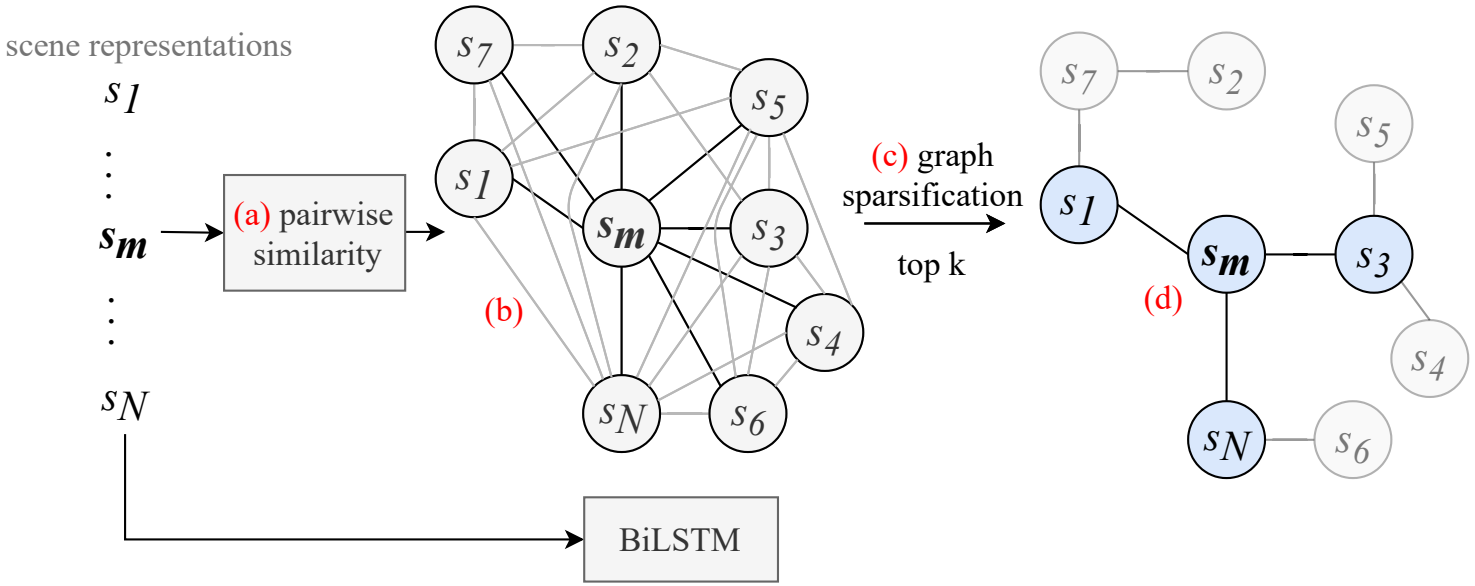
Graph-based Model for TP Identification



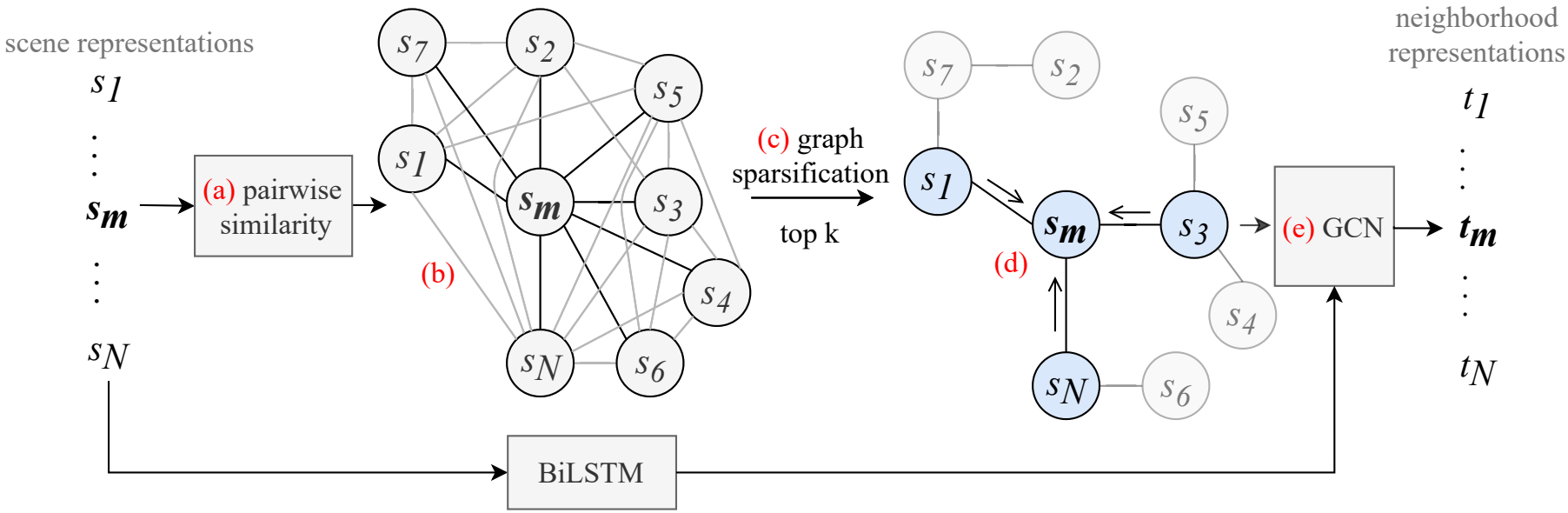
Graph-based Model for TP Identification



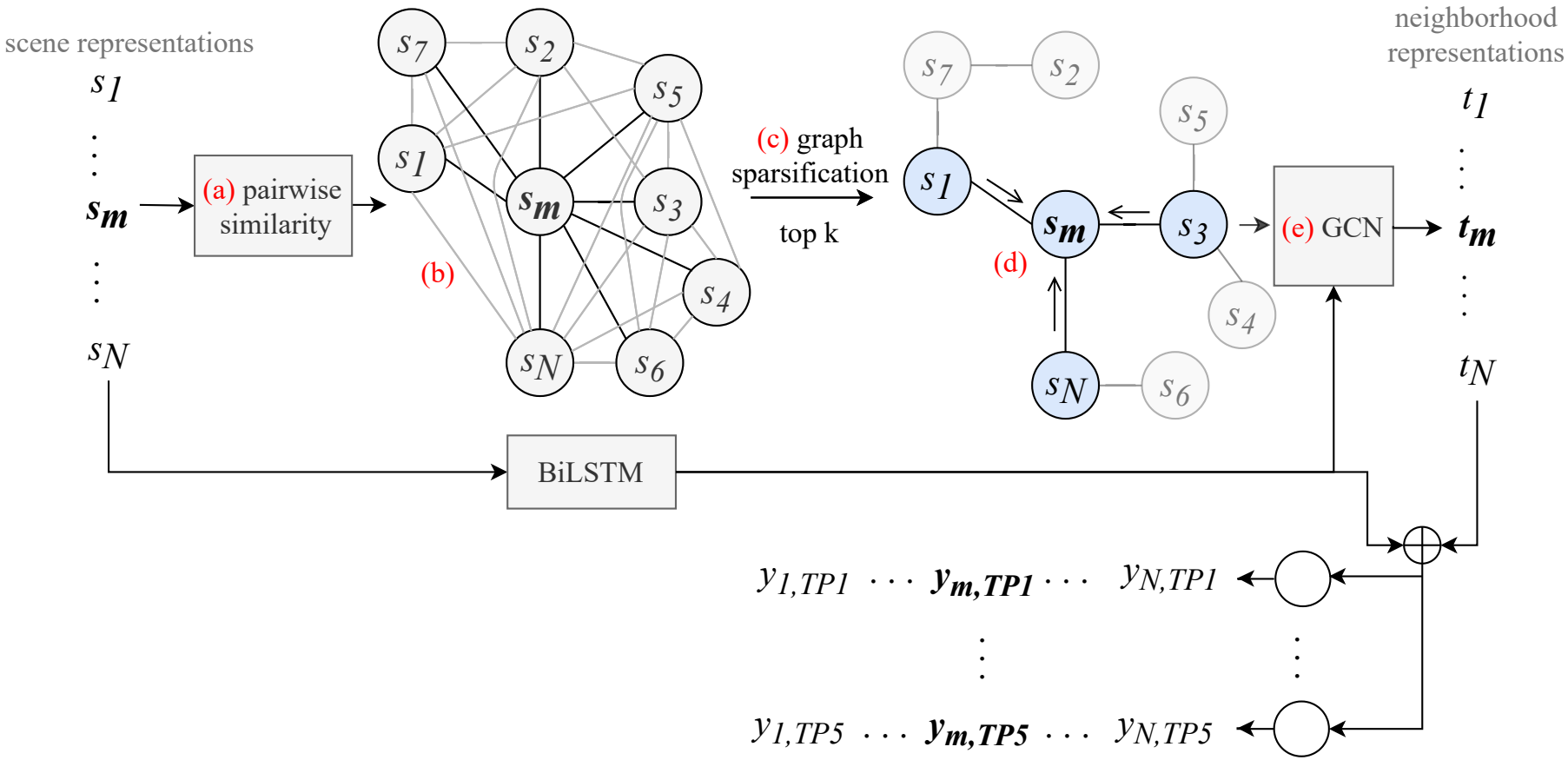
Graph-based Model for TP Identification



Graph-based Model for TP Identification



Graph-based Model for TP Identification



Experimental Setup

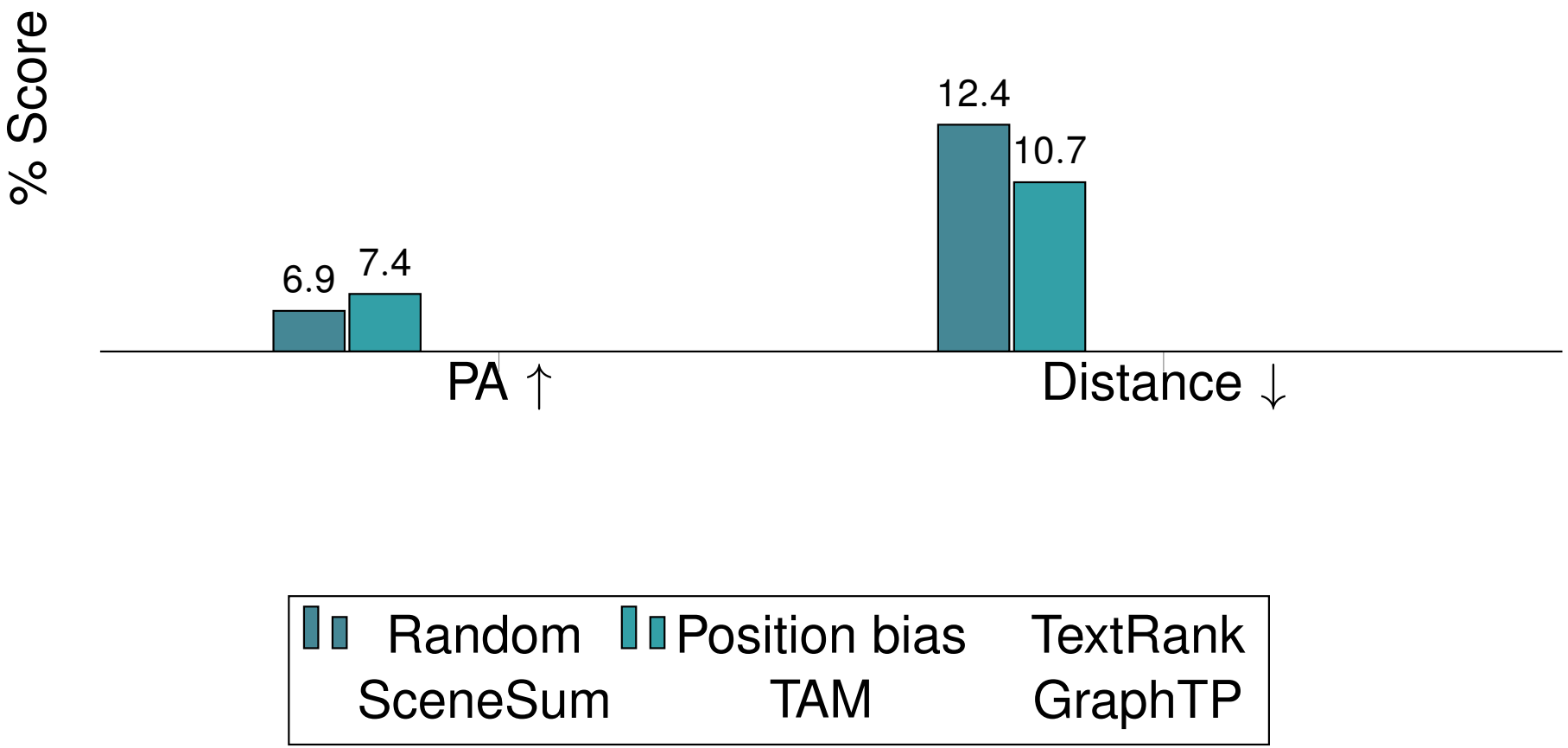
	Train	Test
movies	84	38
scenes	11,320	5,830
TP scenes	1,260	340
vocabulary	37.8k	28.3k

- **TRIPOD** dataset (Papalampidi et al., 2019)
  122 movies; 17,150 scenes; 1,600 TP scenes
- Sparsity: up to **6 neighbors** per scene in movie graph
- Compression rate: 3 scenes per TP, **10–15 minutes** length

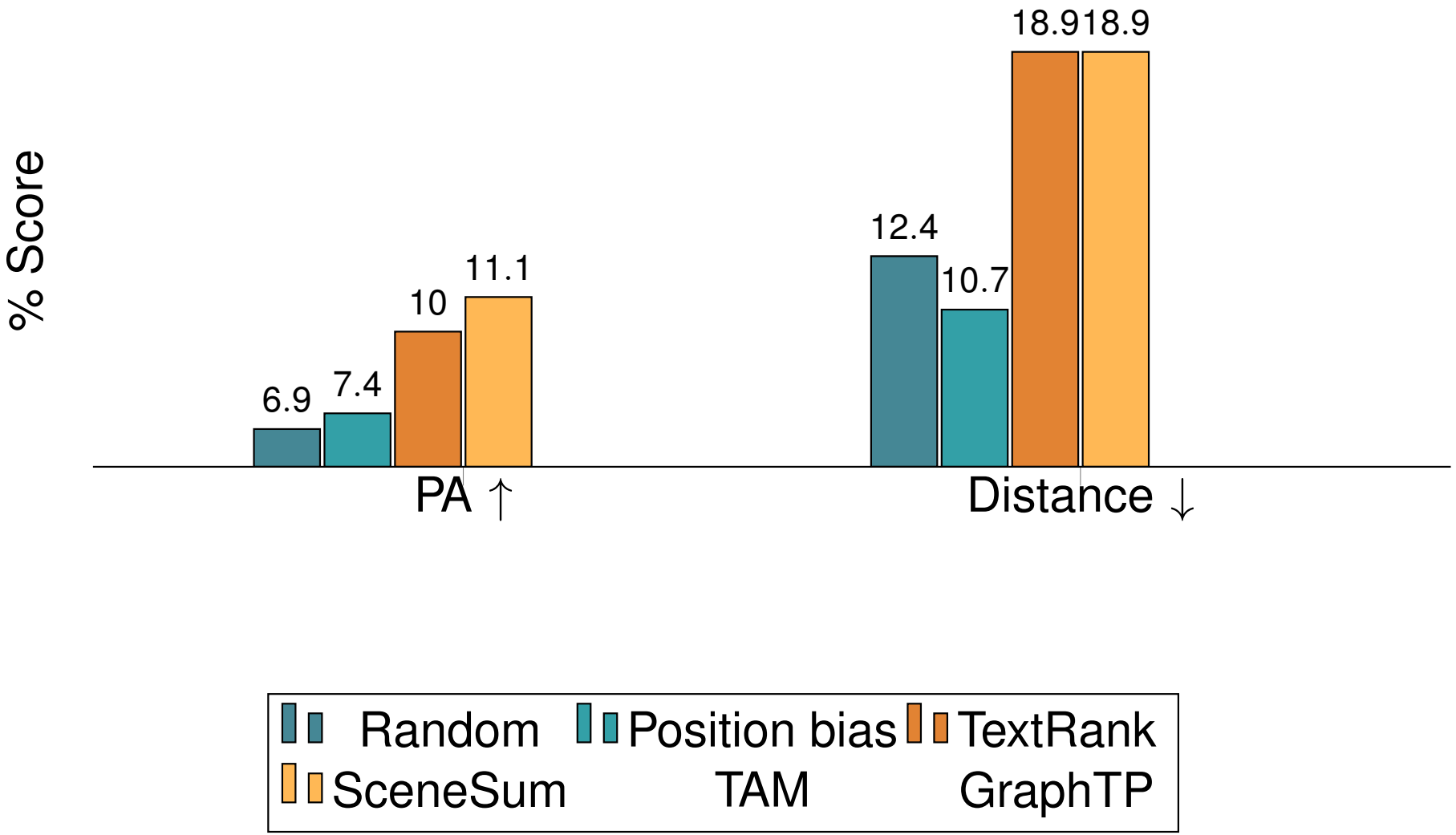
Evaluation Metrics

- **Total Agreement (TA):** the percentage of TP scenes that are correctly identified.
- **Partial Agreement (PA):** percentage of TP scenes for which at least one gold-standard scene is identified.
- **Distance (D):** minimum distance in number of scenes between predicted and gold-standard set of scenes for a given TP, normalized by screenplay length.

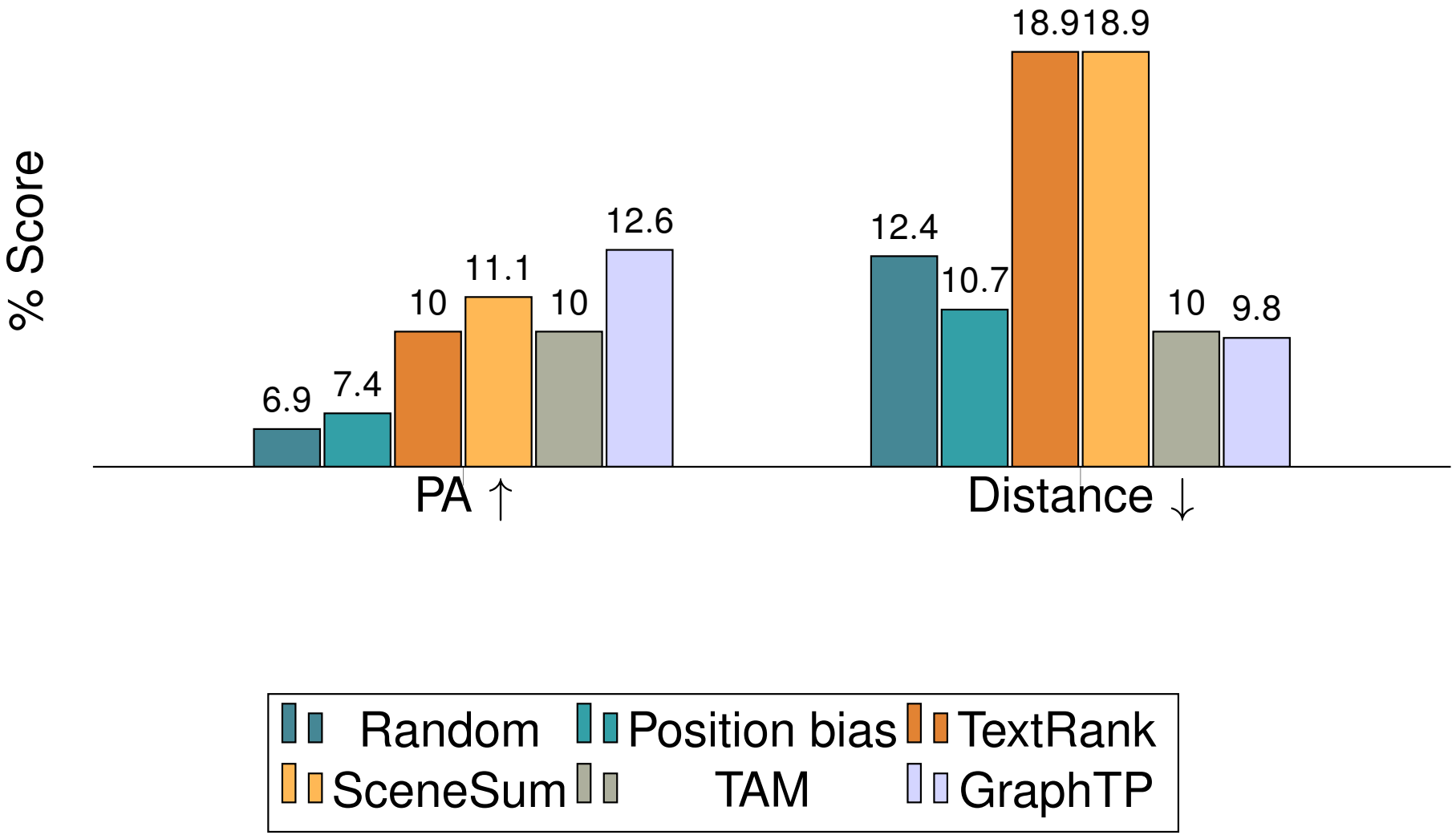
Which Model Identifies TPs Best?



Which Model Identifies TPs Best?

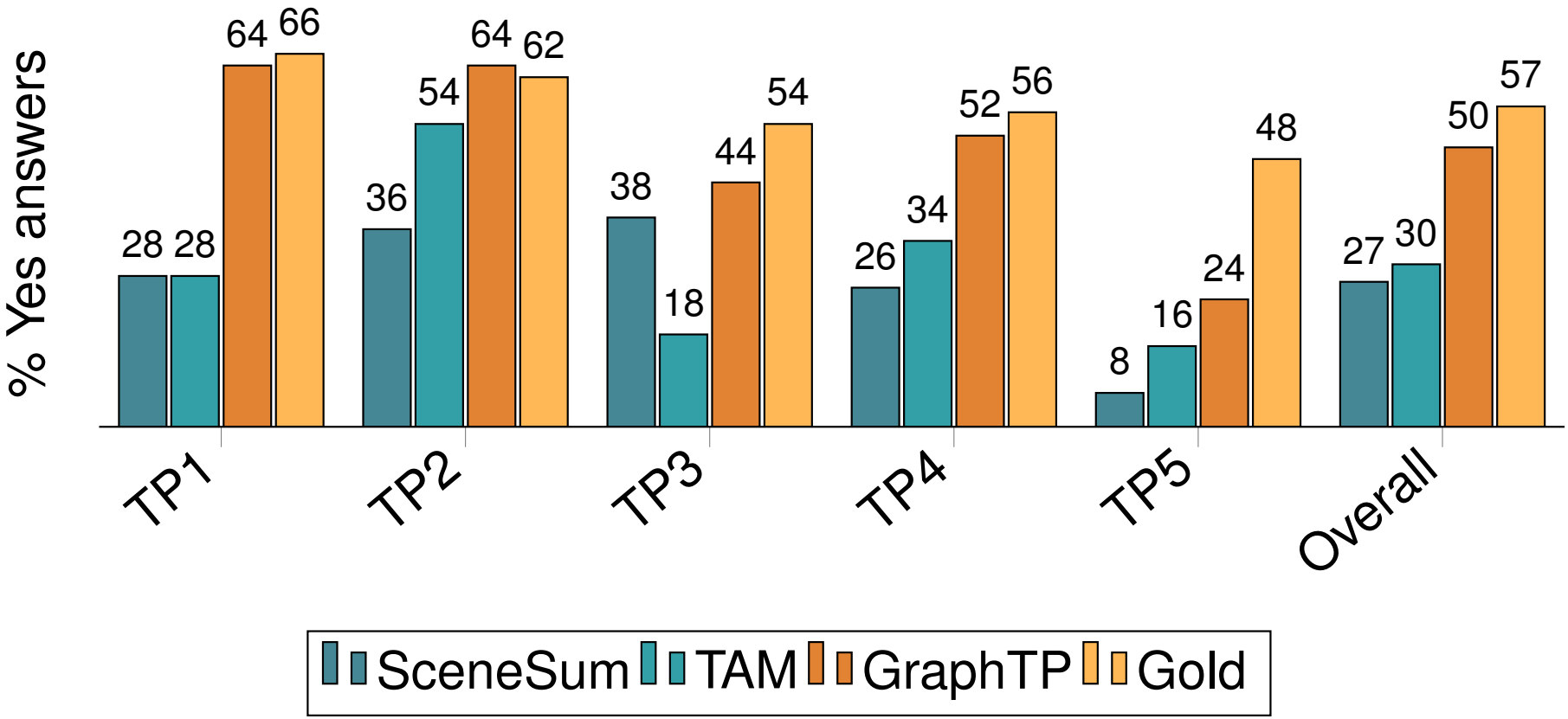


Which Model Identifies TPs Best?



How Good Are the Summaries?

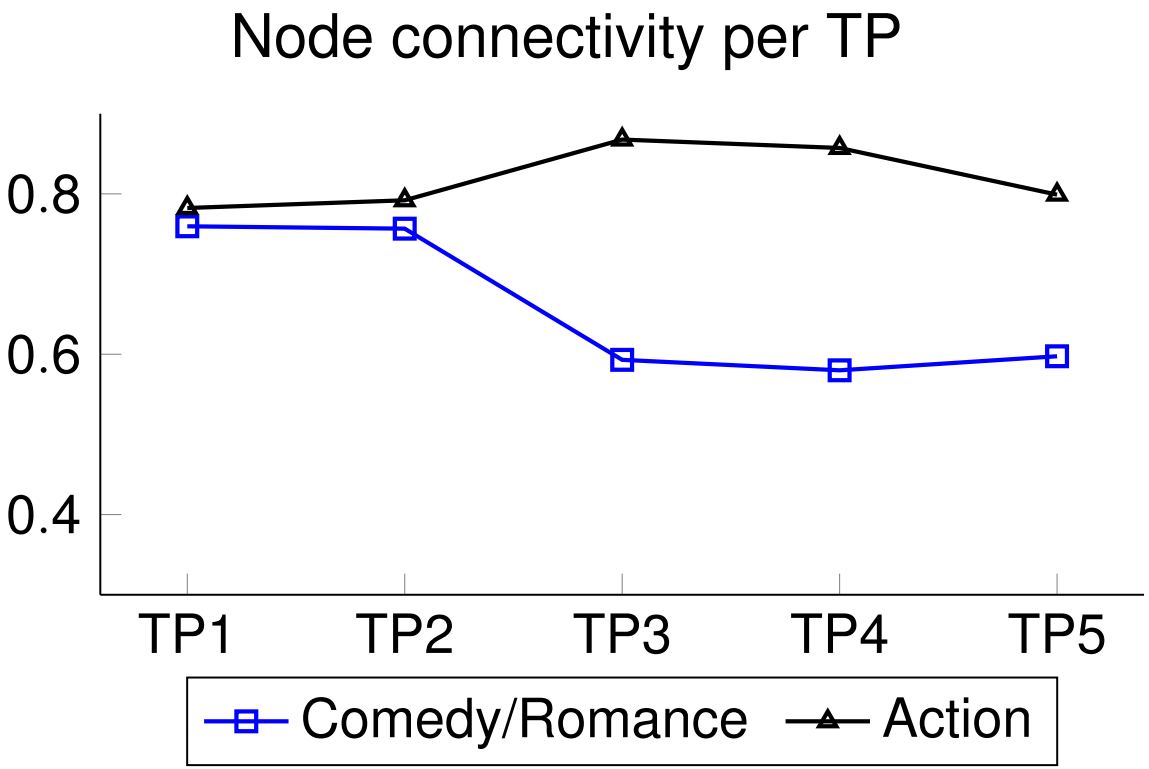
- Do the summaries contain the **key events?**
- Are the summaries overall **informative?**



What Do the Graphs Mean?

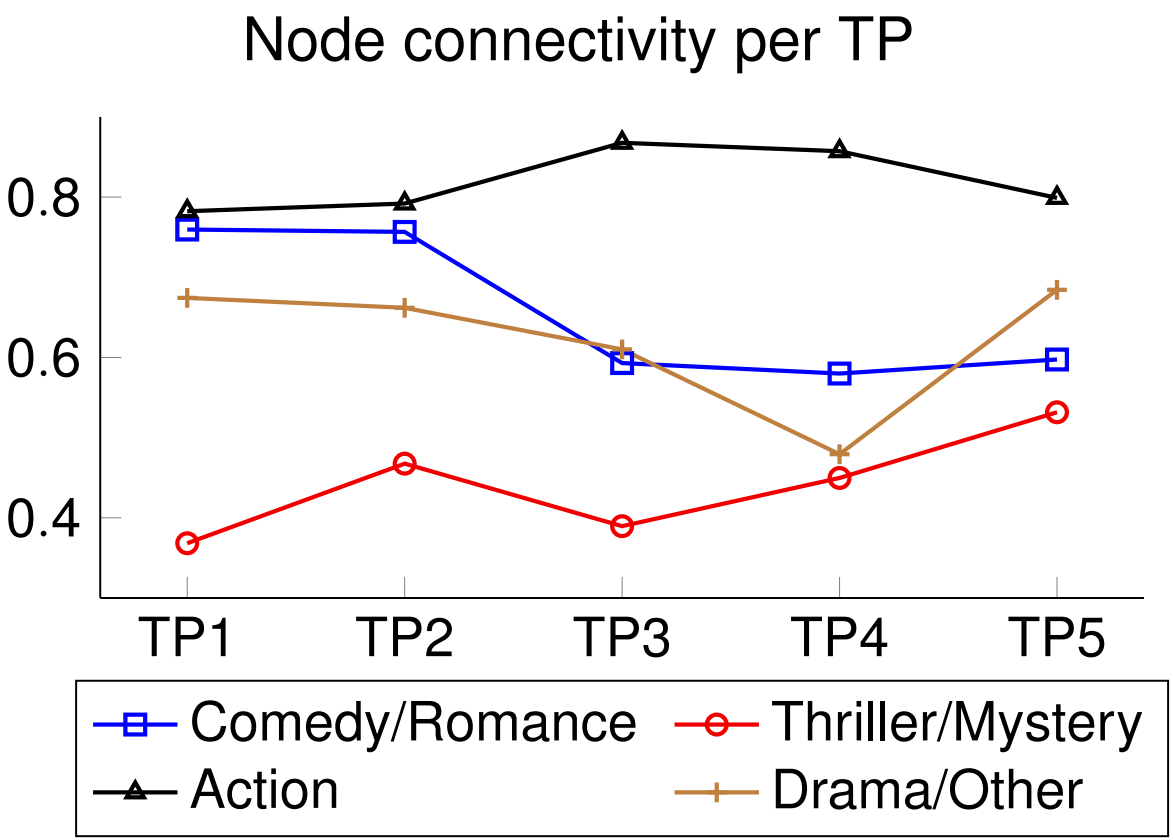
- **Graph Pruning:** only TP scenes and immediate neighbors
- Metric: **node connectivity**
- Movies grouped into **four broad genre categories**

What Do the Graphs Mean?



- **Action, comedies:** high connectivity, more **coherent stories**

What Do the Graphs Mean?

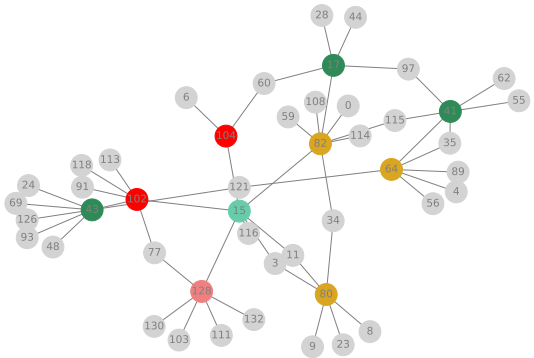


- **Action, comedies:** high connectivity, more **coherent stories**
- **Thrillers, dramas:** low connectivity, more **complex storylines**

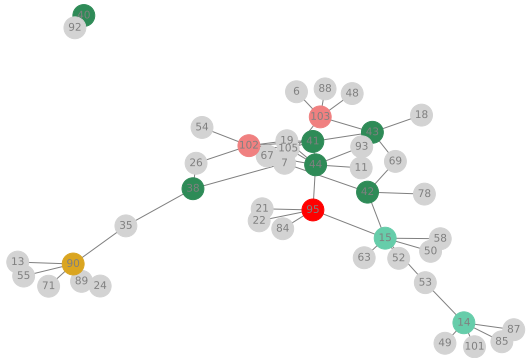
What Do the Graphs Mean?

Comedies and Action Movies

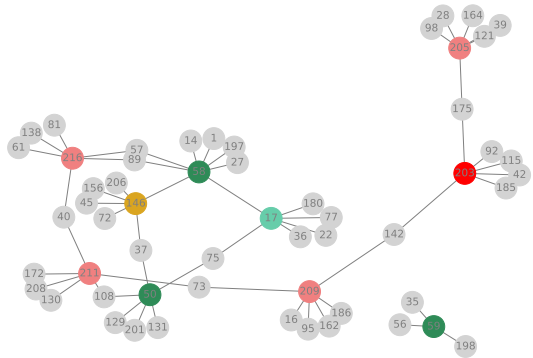
The Wedding Date



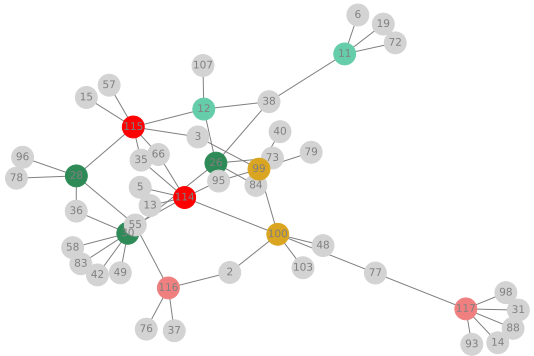
Easy A



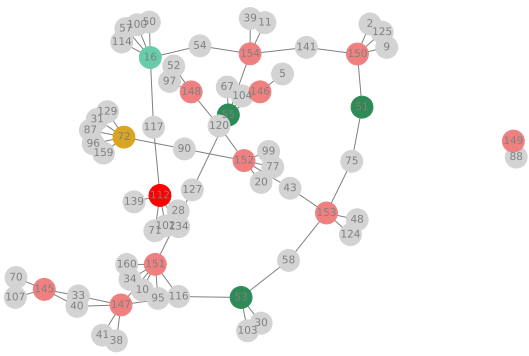
Marley & Me



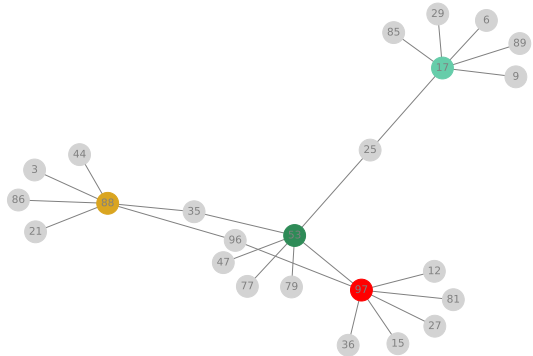
Die Hard



Total Recall



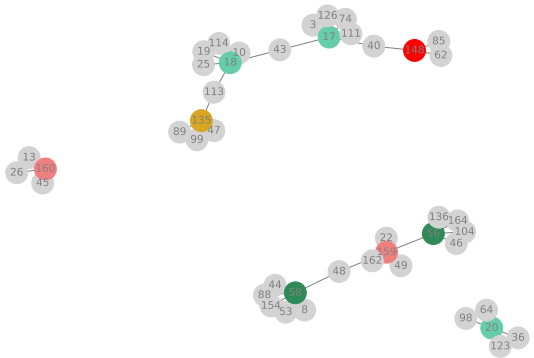
From Russia With Love



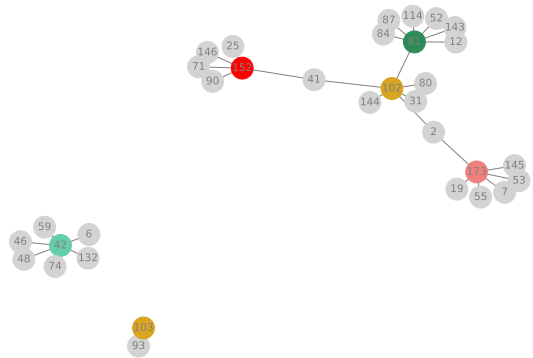
What Do the Graphs Mean?

Dramas and Thrillers

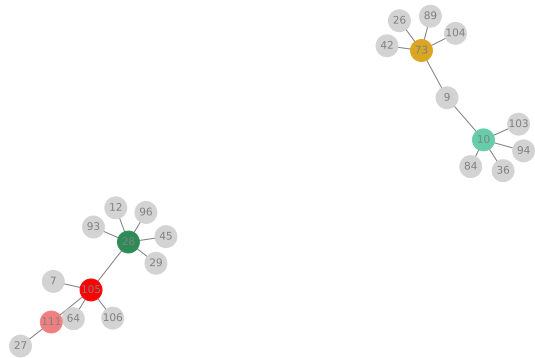
Panic Room



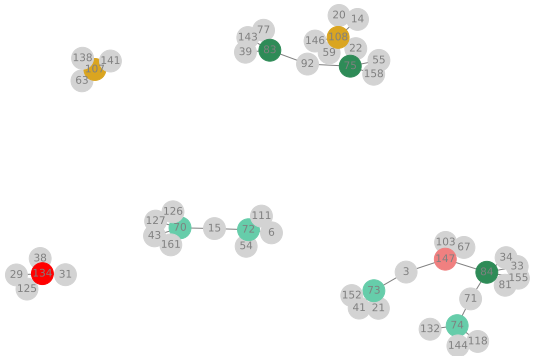
The Thing



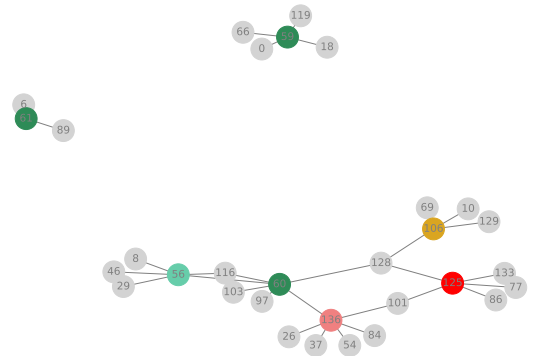
Black Swan



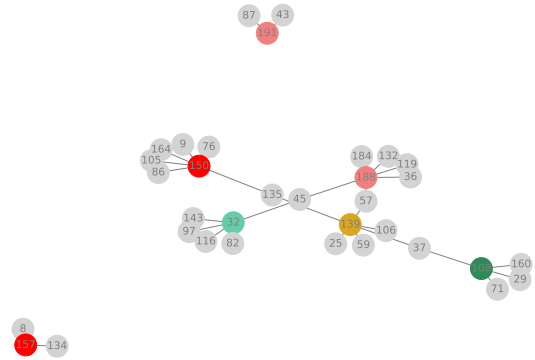
American Beauty



Jane Eyre



Slumdog Millionaire



Conclusions

- Movie summarization via **turning point identification**
- Movies represented by **multimodal sparse graphs**
- Graphs are **interpretable:** model connections between scenes

Current work: **trailer generation:**

- trailers are much shorter ($\approx$2 minutes)
- use **shots** rather than scenes
- no need for screenplays at test time