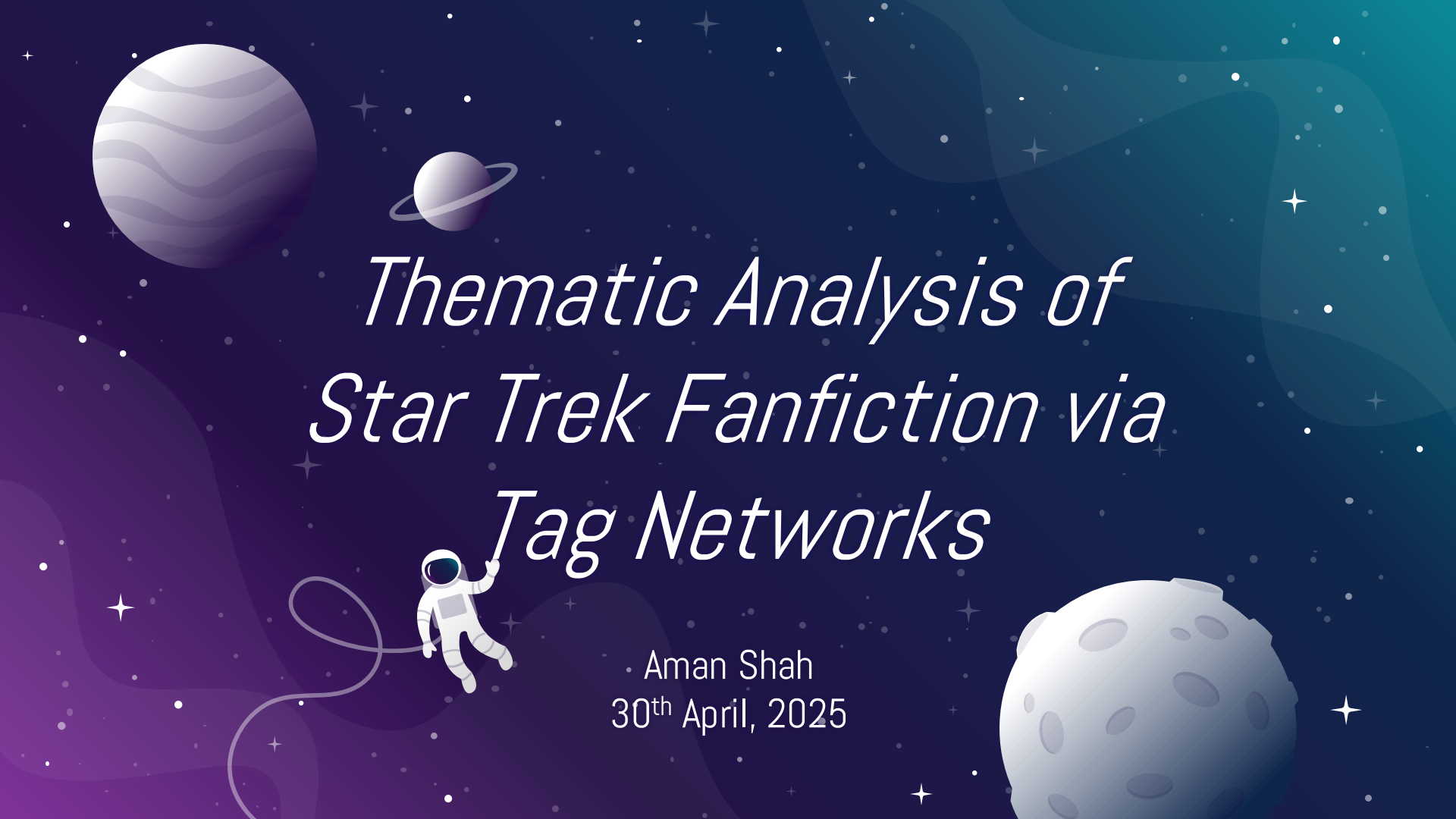


Thematic Analysis of Star Trek Fanfiction via Tag Networks

Aman Shah

30th April, 2025

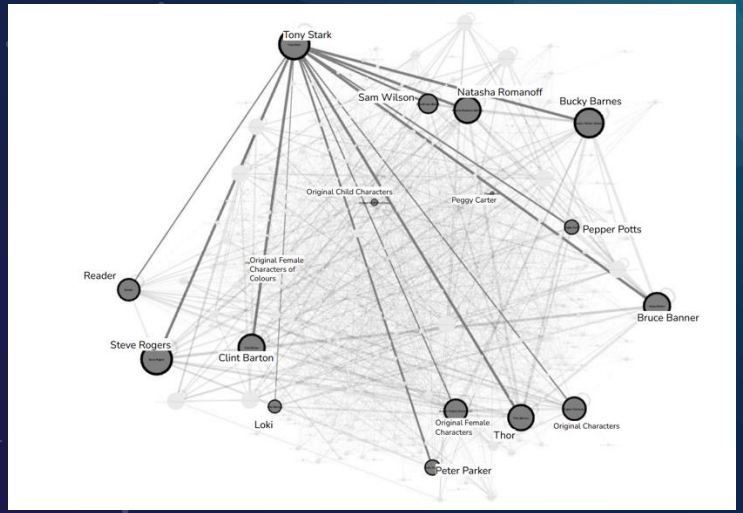
Motivation

Why Study Fanfiction?

- Fandom provides “opportunities for **building mutual understanding**, analyzing meaning, and celebrating other fans’ creations and insights” (ALA)
- Important in upending entrenched notions about **what is normal**, promoting diversity in literature and creative works
- **Active rather than passive** creation of media, promoting imaginative play and learning
- Important to understand what topics fans are writing about to signal **what is important to our society in the current moment**

Past Work

- Archive of our Own is the largest fanfiction website with over 14 million works across 70,000+ fandoms
- Student project on Github wrote a web scraper to get data about stories, their characters, and other metadata
- Also performed initial network analysis to understand **associations between fandoms, between tags, and between characters**



Dataset

Background

- Topic Modeling on the stories is computationally prohibitive, but we want to **understand what people are writing about**
- Stories on **Archive of our Own** have **rich metadata in the form of tags** that could allow for a network-based analysis
- The Star Trek Reboot movies were popular, but are a **small enough dataset** for this type of analysis
- Since there were multiple movies with **several years in between**, we can trace the **change in the fandom** by movie

The Data

Fandom	Star Trek: Alternate Original Series
Total Works	~32,000

Picked the top ~1000 completed English stories for each of the following cases

Date Last Updated	April 22nd, 2013
Reason	Right before the release of *Star Trek: Into Darkness*

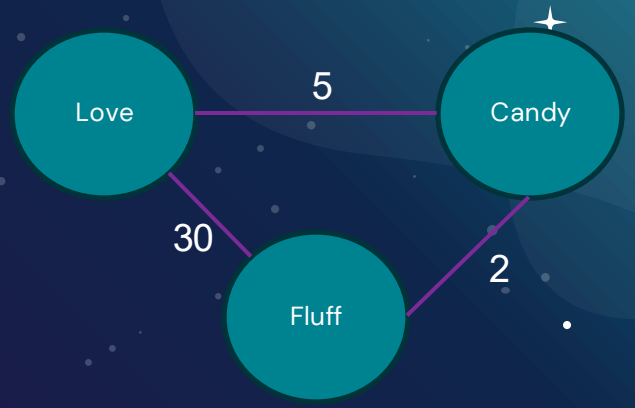
Date Last Updated	July 6th, 2016
Reason	Right before the release of *Star Trek: Beyond*

Date Last Updated	March 31st, 2025
Reason	Most current writing

Used **ChatGPT-4o** to determine **semantically similar tags** and combined them (eg. "Spoack" and "Spock")

The Graphs

Nodes	Unique tags across all stories
Edges	Edge exists if two tags appear on the same story
Weights	Number of stories in common



- For analysis, dropped **hub nodes** like “James T. Kirk” that were connected to a **large majority of other nodes**

Graph 1	Pre-Into Darkness (ID)
Nodes	1469

Graph 2	Pre-Beyond (B)
Nodes	2602

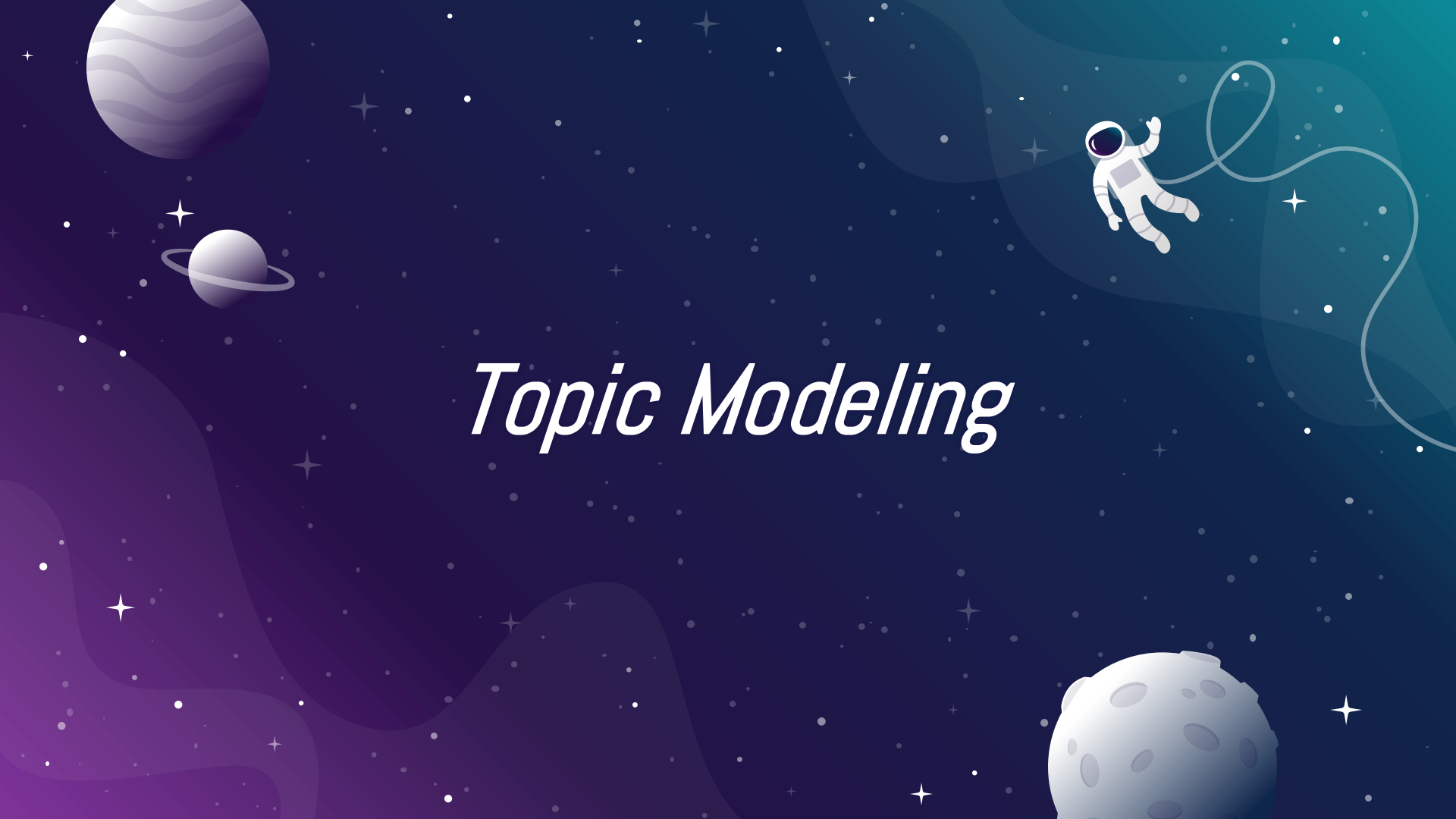
Graph 3	Current (C)
Nodes	3646

Research Questions

- Does analysis of tags **as a network rather than as topics** improve the quality of the results?
- Between modularity and surprise, **which is a better metric to optimize for** in social science settings?

- Have the **topics that fans are writing about changed** as future movies have been released? Can connections be made to the *zeitgeist* of the time?
- What do **larger vs smaller communities** in this network indicate about the fandom?

Topic Modeling



Using BERTopic

- Bidirectional encoder representations from transformers (BERT) is an older language model commonly used in NLP
- BERTopic leverages **BERT embeddings and TF-IDF** to create topic clusters from datasets
- While not as mainstream as unsupervised modeling like K-Means, it represents a **more modern, but still simple way to cluster topics**

Results on G1-3

Topics	1. *Adult content* 2. mccoy, james, leonard, kirk 3. spock, nyota, uhura, spockuhura
Misc. Tags	541 / 1469

Topics	1. kirk, spock, james, mccoy 2. *Adult content* 3. uhura, character, community, au
Misc. Tags	609 / 2602

Topics	1. *Adult content* 2. spock, jim, kirk, james 3. vulcan, trek, star, original
Misc. Tags	900 / 3646

- **Significant portion of tags** were categorized as miscellaneous, meaning that **BERTopic** can't identify niche fandoms/subgroups
- Multiple clusters were created around **character names**, meaning the nuance of certain characters being written about in certain ways was lost
- Adult content was generally grouped into its own category but **often overlapped with things like "love"**, meaning nuance was lost in this as well

Community Detection

Methods

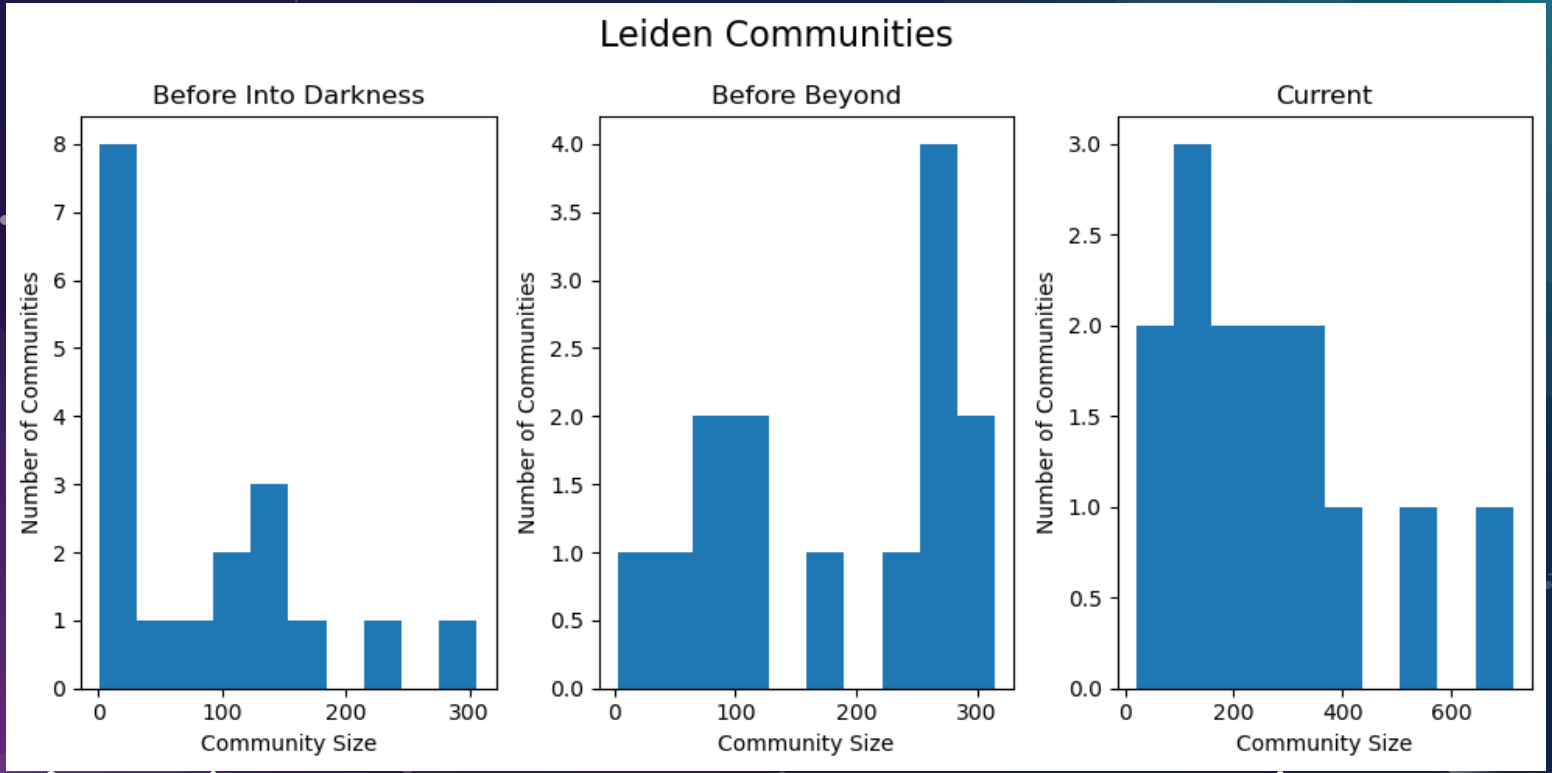
Leiden

- **Implementation of Leiden algorithm** for community detection to understand the types of stories most commonly written
- While this is often an NLP problem using story text, the **rich tag metadata** of these works allows us to treat fanfiction as a social network
- Compare **communities detected** and **strength of communities** across the multiple time periods discussed

Surprise

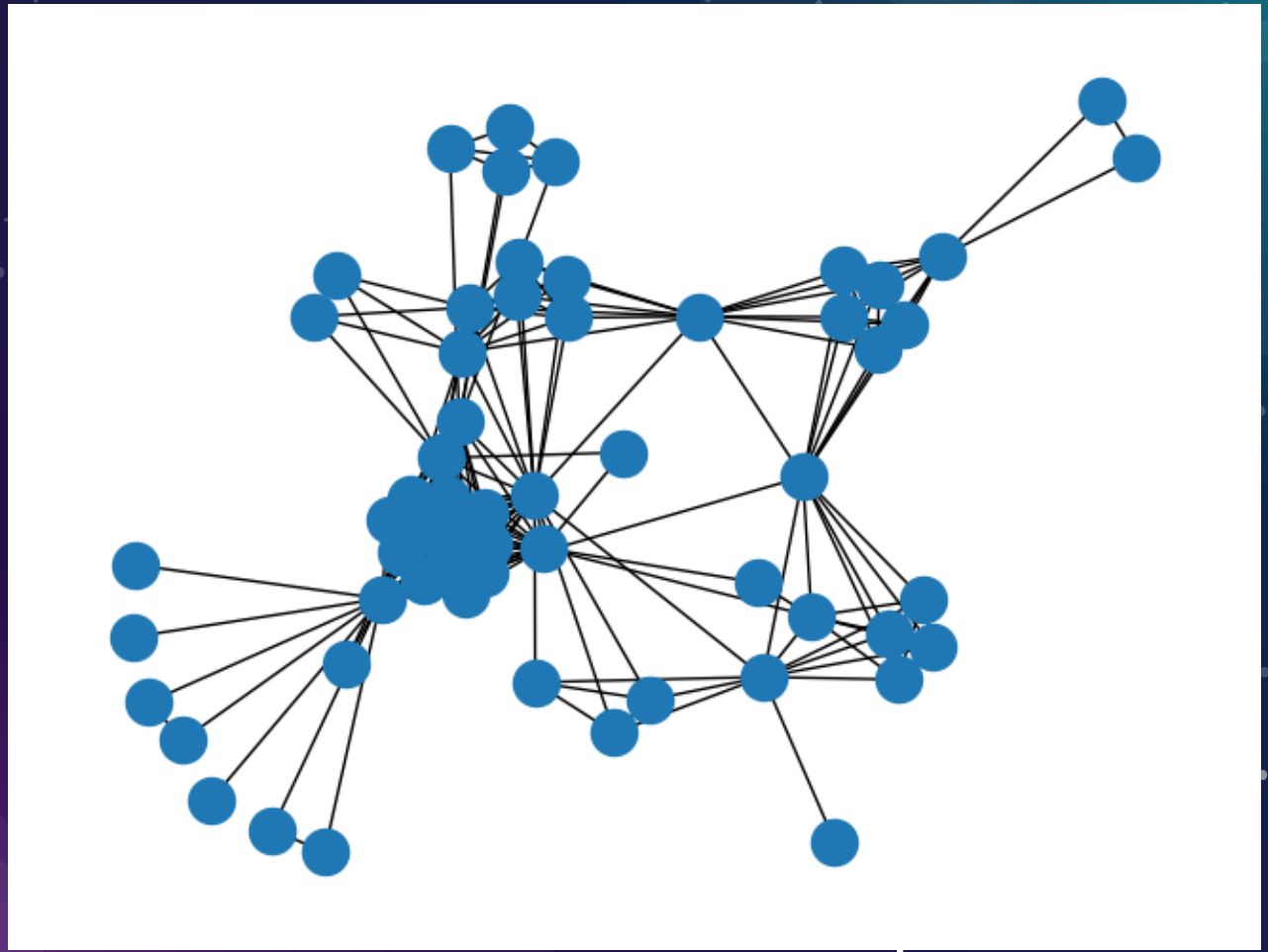
- Aldecoa and Marín propose for optimizing Surprise, which **measures the "surprise" (improbability) of finding by chance a partition in a random graph with the observed intracommunity links in the network**
- Greedy algorithm that moves nodes from one community to another to improve Surprise
- Better resolution than modularity

- Used **ChatGPT-4o** to obtain general themes and topics for each identified community


Leiden Communities
Before Into Darkness
Before Beyond
Current
Number of Communities
Community Size
0
100
200
300
400
600
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
1
2
3
4
5
6
7
8

Complex Character Studies and Pre-Canon Development

Graph 1



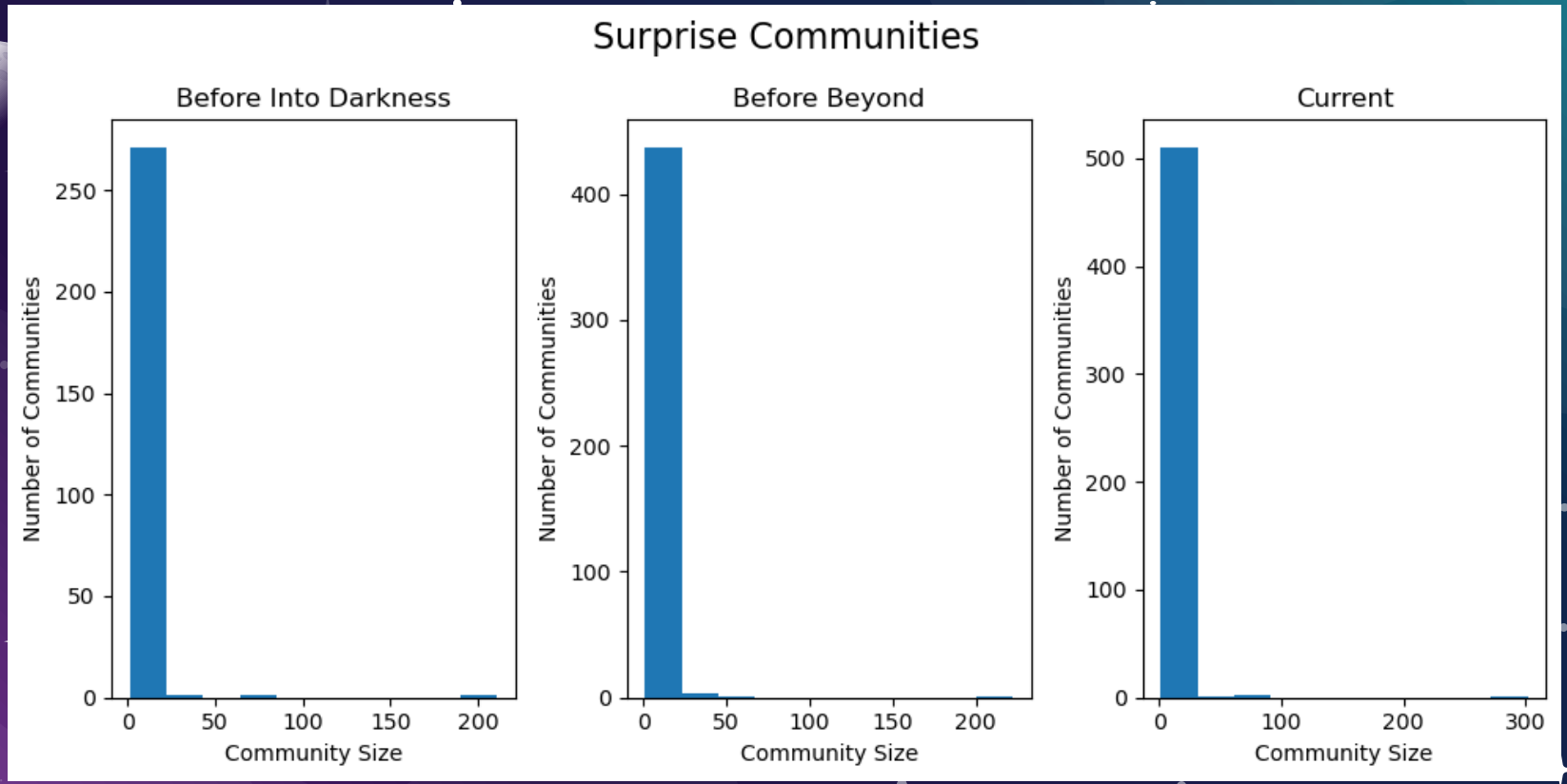
Leiden Results on G1-3

Topics	• Emotional Trauma and Relationship Complexity • Classic Tropes and Feel-Good Pairings

Topics	• Star Trek Romance with Fairytale and Royal Elements • Star Trek Slice of Life and Emotional Healing

Topics	• Humor / Crack, Alternate Universe, Fluff • Coming of Age / School, Trauma / Recovery, Mental Health, Bonding / Soulmates

- ★ Topics identified were **more complex and nuanced,** focused on specific types of stories and settings
- ★ Adult content was **far more common** in the last graph**, signifying a shift over time**
- ★ Leiden was unable to identify **highly specific themes** on the largest graph, showing problems with **Leiden's resolution**


Surprise Communities
Before Into Darkness
Number of Communities
Community Size
0
50
100
150
200
250
Before Beyond
Number of Communities
Community Size
0
100
200
300
400
Current
Number of Communities
Community Size
0
100
200
300
400
500

Surprise Results on G1-3

Misc. Themes	156 / 274
Topics	*Adult content* Mirror Universe

Misc. Themes	248 / 443
Topics	*Adult content* Alternate Universe

Misc. Themes	280 / 514
Topics	Hurt/Comfort *Adult contet*

- ★ More than **half of all communities for** each graph were determined to have “miscellaneous theme”, likely **due to the small size of communities**
- ★ While hypothesis was that the higher resolution would be beneficial, 1 and 2-node communities are **not helpful for analysis**
- ★ Only discernible themes are **more general and very similar to already existing tags**


Conclusions
+
Future Work

Comparing the Methods

- After dropping hub nodes and combining similar tags, **Leiden was far more appropriate for thematic analysis**
- Surprise led to **a massive number of 1 and 2-node communities**
- Tag networks were more effective than BERTopic as they allowed for **complex story ideas** that cannot be accurately captured in a single tag
- Over time, **adult content was far more pervasive** in the fandom, but across all time periods, **nearly 100% of content had romance involved**

Future Work

- Due to computational limitations, **only top ~1000 works were considered,** so all niche tags were not included
- Hierarchical agglomerative clustering on the Surprise communities **could help solve the issues with very small communities**
- **Quantitative methods** to evaluate the quality of these communities as there is no ground-truth for this data


Questions?