Analyzing Identity Clusters behind #sahm TikTok Creators: A Study on the Creator's Use of Hashtags and Their Suggested Niche Identities

Yolanda Zhang
Wellesley College | Data Science Major Capstone

Introduction

Niche hashtags on TikTok are hyper-specific hashtags used by creators to categorize contents that would align themselves with specific identity communities (e.g., #sahmlife, #momsoftiktok, #breastfeedingjourney, #adhdmom). While some niche hashtags focus on relatable moments or personal storytelling (such as #momfail, #toddlermeltdown, #mentalmoments), others signal more stable identity positions tied to roles, values, or communities (e.g., #christianmom, #gentleparenting, #crunchymom).

As the hashtag #sahm (abbreviation for the term *Stay-At-Home-Mom*) has become increasingly popular on TikTok, this study investigates how #sahm creators construct and maintain distinct, personal identities through their use of niche hashtags.

Research Question

How do #sahm creators form and maintain niche identities on TikTok through their hashtag patterns?

Data & Methods

The data contains #sahm TikTok posts collected between 2018 to 2025, with all identifiable users' information removed.

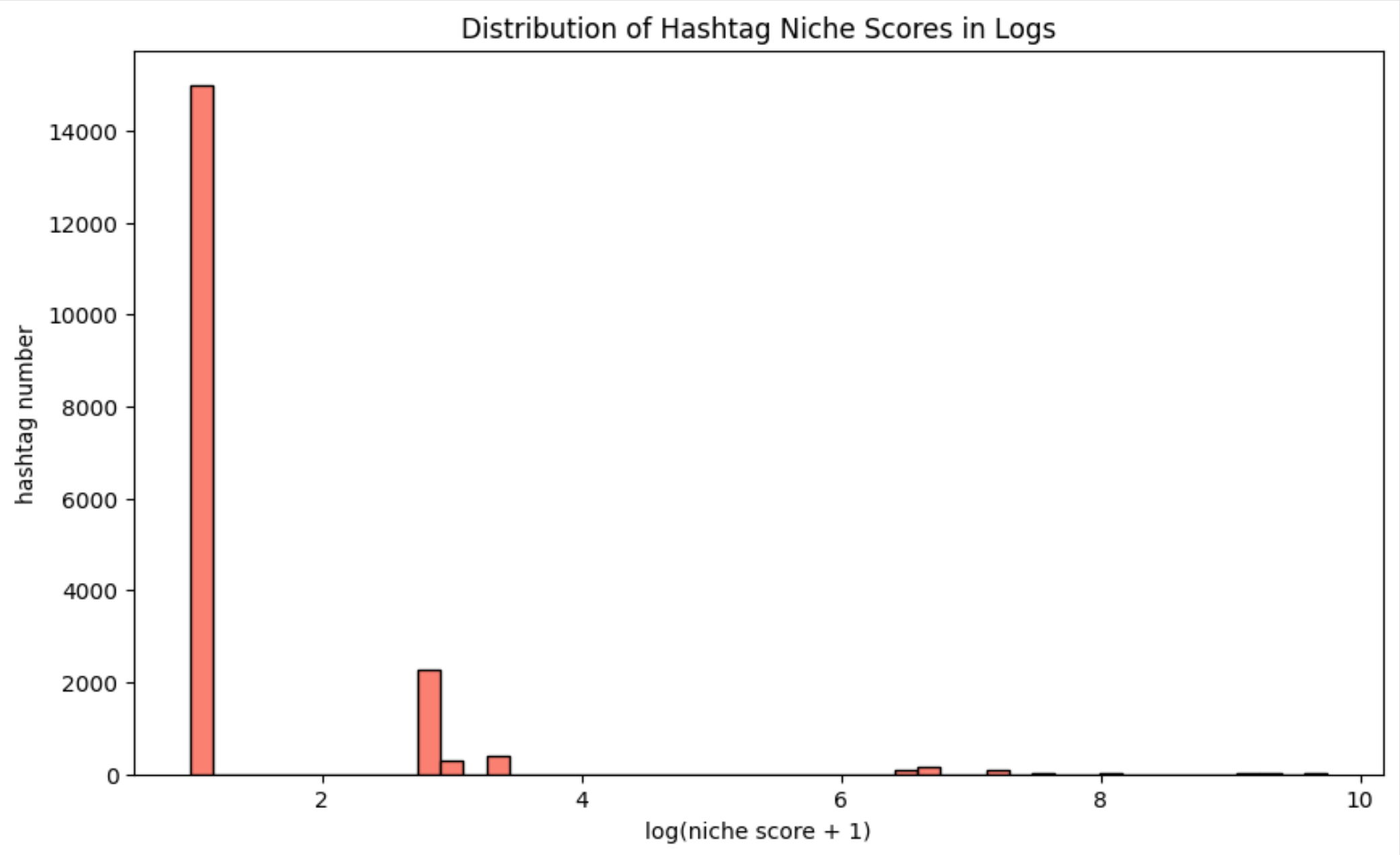
Definitions

- **Hashtag Co-occurrence Network**: identify and count the number of times each pair of hashtags has occurred together in a post
- **Louvain Method**: group nodes into communities by measuring how densely connected nodes are within groups compared to across groups
- **Niche Score**: measure how strongly a creator uses hyper-specific (niche) hashtags in their posts, with higher niche scores indicating more cluster-specific (niche) usage
- **Hashtag Communicative Function**: the role a hashtag plays in communication (e.g. it labels the topic, expresses a stance or identity, or connects the post to a specific community or trend)
- **Entropy Score**: measure how spread out or consistent a creator's hashtag usage is, with lower scores indicating the creator uses similar niche hashtags (more consistent identity)

Niche Hashtag Clusters Construction

- Build a **hashtag co-occurrence network table** from posts (11.7M hashtag pairs), keeping only pairs appearing **>= 50**
- Apply **Louvain community-detection algorithm** to the co-occurrence network to group hashtags into **14 identity-based clusters**
- Compute a **niche score** for each hashtag: within-cluster popularity (count / total count in the cluster) ÷ between-cluster popularity (count / total count)
- Draw a **distribution** of all hashtags' niche scores (log-scaled for interpretability) to select a threshold above which the hashtags would be "niche" (see Fig. 1)
- Set **niche cut-off** = **x ≥ 3**
- Create an indicator variable that identifies **864 hashtags** above the niche cut-off
- Remove overly broad or extremely rare hashtags from the 864 identified hashtags → end up with **480 niche hashtags**

Figure 1. Hashtag Niche Score Distribution (Log)

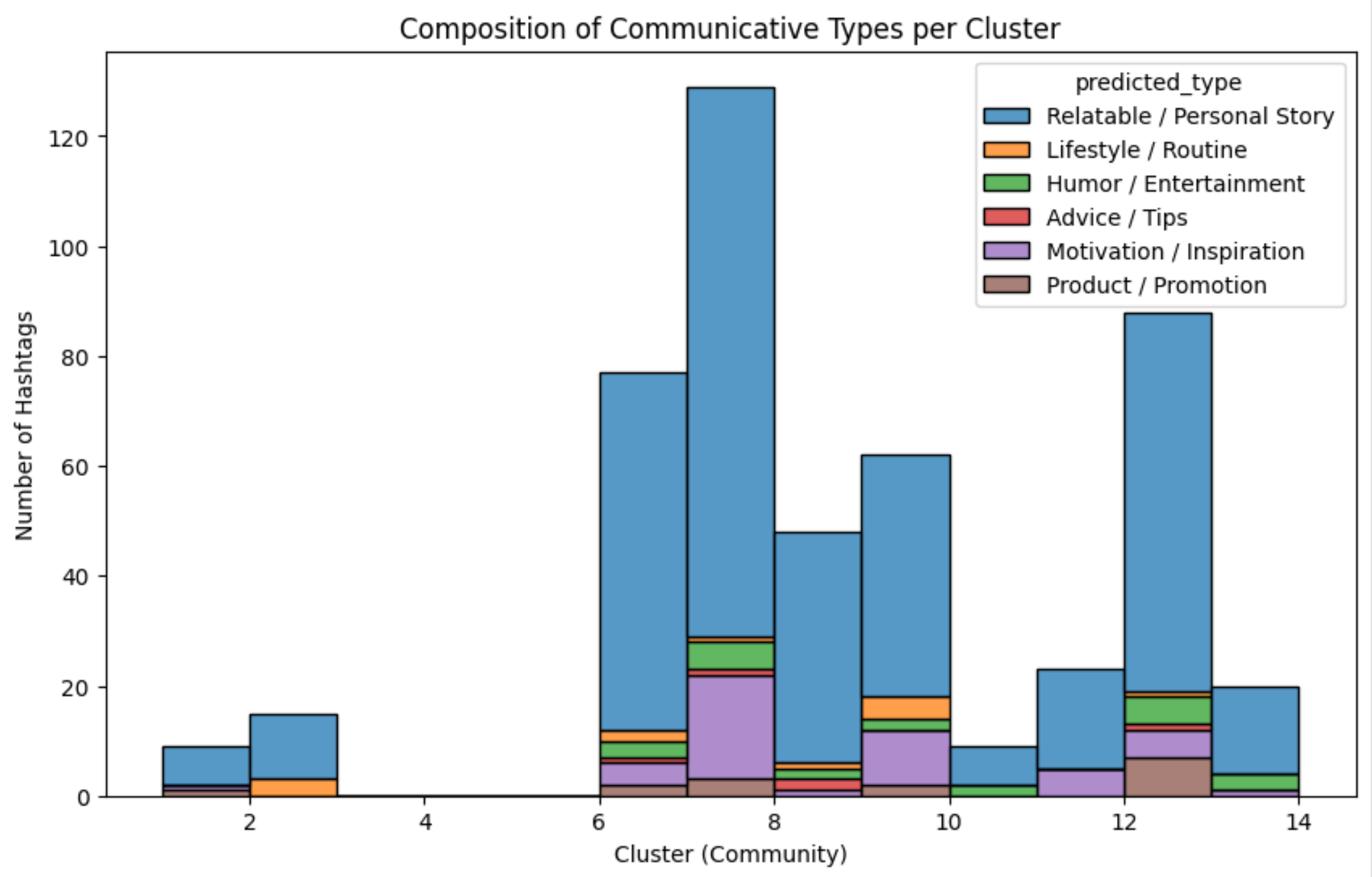


- Use **zero-shot learning** (classify new categories without any training examples) to categorize each niche hashtag into one of the **six communicative functions** (see Fig. 2)

Communicative Function Examples:

- Relatable / Personal Story: #sisterlove
- Dvice / Tips: #longdistance
- Lifestyle / Routine: #spreadlovenothate
- Humor / Entertainment: #ketolifestyle

Figure 2. Hashtag Cluster Communicative Type



Creator-level Identity Construction

- Extract niche hashtags (appearing in ~1% of posts) and assign them to each post
- Build a **post X niche-hashtag matrix** indicating which identities appear in each post
- Aggregate to the creator level to compute: **counts** of each niche identity used and **proportion** of their posts containing each identity

Identity Consistency

- Compute **a normalized entropy score** using the Shannon entropy formula: $\mathbf{H = -\sum p_i \log(p_i)}$, where p_i is the proportion of posts tagged with identity *i* with **higher H = more diverse identity usage**
- Assign **entropy score** to each creator to measure how consistently they position themselves across identities
- Classify creators by their entropy score and assign their consistency: **consistent** (< 0.25), **mixed** (0.25–0.55), **fluid** (> 0.55)

Creator Identity Clustering

- Apply **PCA** (a measure to capture the most **important patterns or variation** in the data) to reduce the high-dimensional identity space
- Use **k-means** (k = 8, the optimal cluster number chosen by Elbow Method: pick the point where the decrease levels off) to identify **major creator identity clusters**
- Each of the 8 clusters represents a group of creators who use **similar mixtures** of niche hashtags
- Compute creators' **principal component 1** (direction that gives the most variation) and **principal component 2** (direction that gives the second-most variation) from **PCA**
- Visualize cluster groups in PCA space and **identify separated clusters**: creators who rely on a distinct and rare set of **niche identity hashtags** that are **not shared** with other creator communities

Identity Network Structure

- Compute **cluster adjacency & nearest neighbors** to detect overlap between clusters: only **~2.25%** of creators show overlapping identity profiles
- Calculate **identity co-occurrence** across creators and remove rare pairs (< 20 creators)
- Identify **bridge identities** (identities that co-occur widely and connect multiple clusters)

Within-Cluster Dynamics

- Measure **identity switching** over time
- A creator is **persistent** if switch rate < 0.3, and dominant cluster > 70% of their posts
- Draw **distributions** of within-creator consistency and fluidity across each cluster (see Fig. 3 & 4)

Final Data Frames:

- **Post DataFrame**: contains IDs for creators and posts, list of hashtags, and list of niche hashtags
- **Hashtag DataFrame**: contains a binary indicator that identifies if the hashtag is niche and cluster assignments for all hashtags
- **Creator-level DataFrame**: contains a vector describing which niche hashtags the creator has used and cluster label for each creator

Figure 3. Within-Creator Consistency

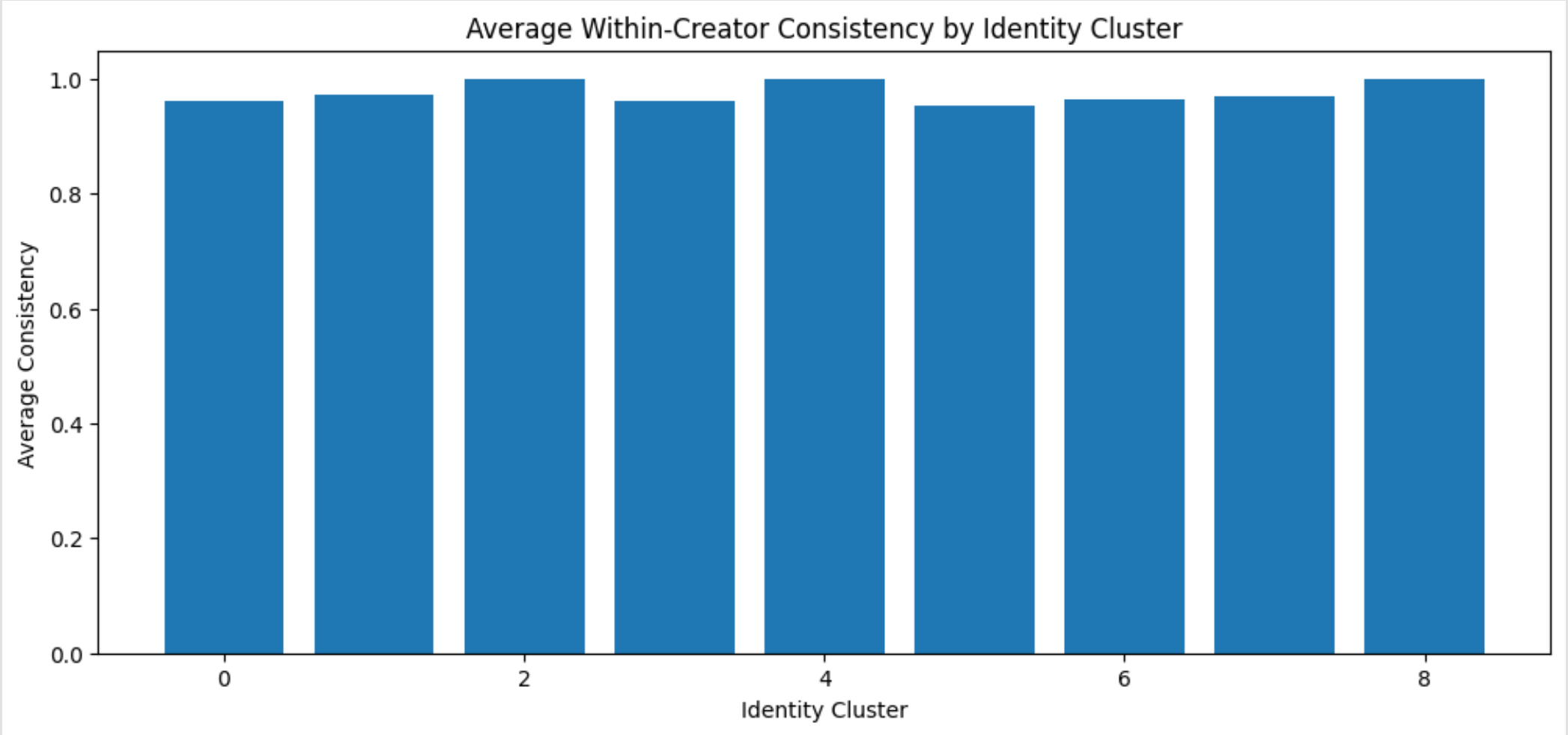
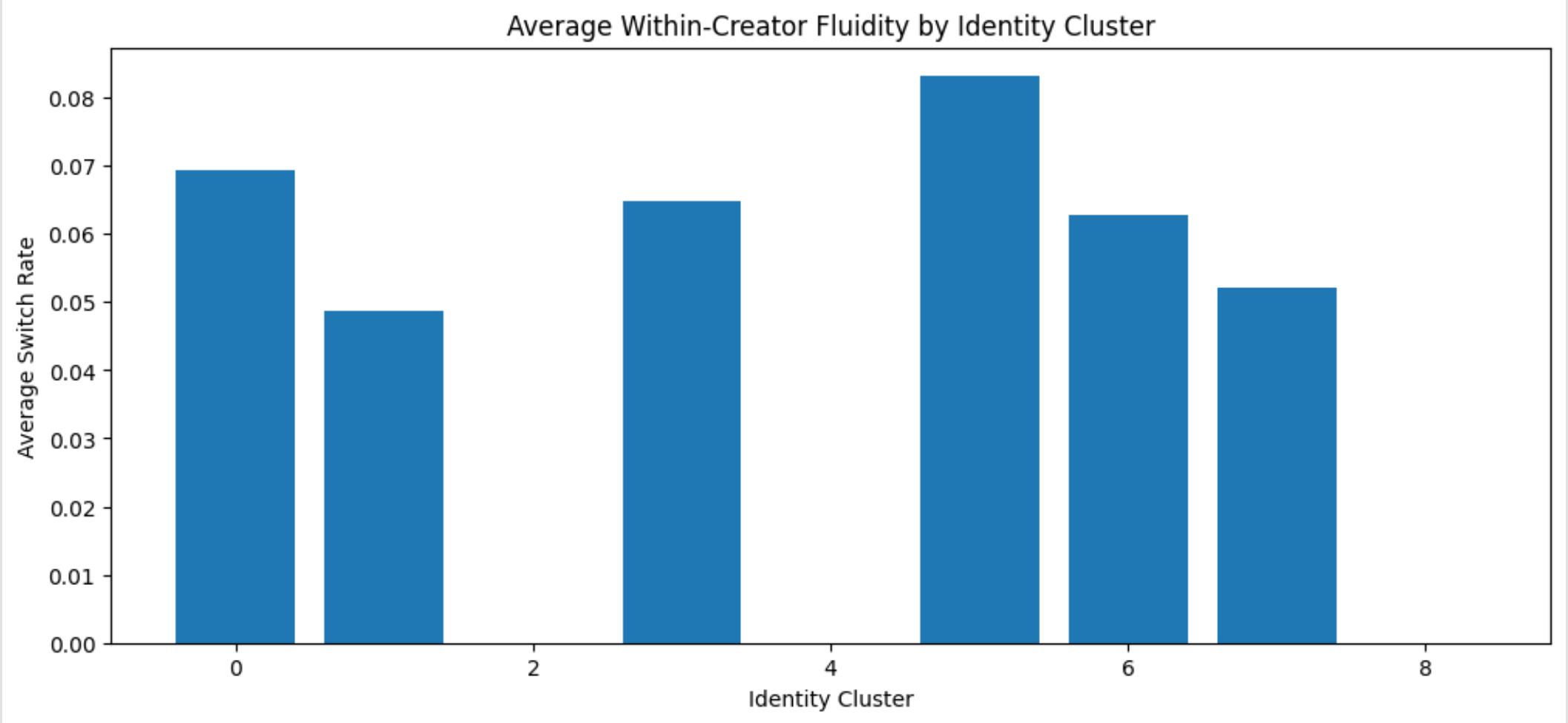


Figure 4. Within-Creator Fluidity



Results

- **Relatable / Personal Story** is the dominant communicative type across niche hashtag clusters, which indicates that creators primarily use niche hashtags to frame content through personal and relational narratives.
- About **99.97%** of creators maintain a consistent identity style, with **0.02%** of mixed style creators, and **0.01%** of fluid creators, which indicates minimal identity switching across posts.
- Cluster 1 (six creators) and Cluster 2 (one creator) **do not share** identity patterns with any other creator identity clusters (macro-level). In contrast, all remaining clusters exhibited **overlapping identity patterns**, suggesting shared identity signals across creators.
- Cluster 2 (size of 27), Cluster 4 (size of 1), and Cluster 8 (size of 3) are **perfectly dominant** with an average dominance proportion of 1.0 and a switch rate of 0.0.
- Cluster 5 has the **highest** switch rate of 8.3%.
- The top 3 identities in Cluster 2 are: "girlsthatwearjewelry", "stainlesssteeljewelry", and "goldplatedjewelry"
- The top 3 identities in Cluster 4 are: "gratefulthankfulblessed🙏🏻", "mamaof3💙💜💙", and "wifey💍"
- The top 3 identities in Cluster 8 are: "mamabearof4", "mammabearof4", and "hallowsmonth"
- Top 5 co-occurring identity pairs & Top 5 bridge identities:

Identity 1	Identity 2	Count
acotar	sarahjmaas	54
reader	readersoftiktok	44
acotar	fourthwing	37
bookrecs	romancebooks	35
acotar	sjm	34

Bridge Identity	Count
ex	974
bookrecs	861
acotar	849
psychic	845
psychicreading	826

Conclusion

- #sahm creators form **highly stable niche identities** [99.97% maintain a consistent persona]
- Many clusters show **perfect identity dominance** [creators post exclusively within one niche]
- Cluster 1 & 2 are **uniquely distinct** and have no shared identity pattern
- Cluster 5 show **notable cross-identity** behavior [creator navigate several adjacent niches (identity bridges)]
- Highly specialized niche identities in Cluster 2, 4, and 8 reveal **micro-communities**
- There's **shared / overlapping communities** within #sahm indicated by the co-occurring identity pairs

In short, most #sahm creators **don't constantly change** their only persona. They stay with **one niche** and use hashtags that match that niche, which builds **small, stable communities** with their own styles and identities on TikTok.

Acknowledgements

Thank you to Professor Eni Mustafaraj from the Data Science Department of Wellesley College of the help along the way. The Github code has received help from GenAI.