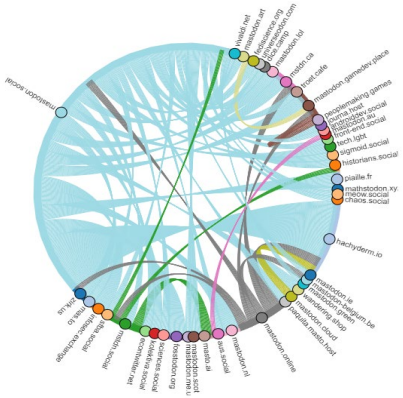


Three-Layer Structural Dynamic Governance Framework for Mastodon's Decentralization
—— An Empirical Study Based on Cross-Instance Interactions

Data:Shen Tianyu; RQ1,2: Xing Hanbo, Luo Riqi; RQ3:Li Jianing

Motivation

Core Background

As a benchmark platform for decentralized social networks, Mastodon's ecological health depends on two key dimensions:

1. Frequency and quality of cross-instance interactions
2. Decentralized balance of the overall network

Gaps in Existing Research

Three unresolved issues exist in the current field:

1. How instance size quantitively affects Cross-Instance Interaction Ratio (CIIR) remains unclear

2. The driving mechanism of factors like language and topics on cross-instance interactions lacks systematic analysis

3. A multi-dimensional decentralization evaluation system is absent

Research Overview

RQ1: Behavioral Motivation Layer: How does instance size affect cross-instance interaction?

RQ2: Semantic Structure Layer: How Languages and Topics Jointly Shape the Semantic Community Structure

RQ3: System Architecture Layer: Multidimensional Centrality and Optimal Scale Window

Data-The Build of The Decentralized Dataset On Mastodon

Observation of the current dataset

The dataset comes from：

https://zenodo.org/records/14869106

The dataset includes：

1. livefeeds.json

It is a snapshot collection of the Mastodon platform, containing 1,361,708 posts.

It mainly consists of four parts: identification (sid), time and interaction counts, account information, and topic information.

2. boostersfavrourites.json

It stores interaction data of boosts and favourites.

It mainly includes three parts: association with specific sid, boost and favourite counts, and core account information related to interactions.

3.reply.json

It stores reply interaction data.

It mainly includes three parts: association with specific sid, initiating account, and replied account.

Dataset of Mastodon Toots (collected by FediLive)

FDUDataNET

This snapshot of Mastodon is captured using FediLive, covering publicly visible activities across the entire platform over a 13-day period, from Nov. 22 to Dec. 4, 2024 (UTC+0). During this period, FediLive collected 1,361,708 original posts and 2,628,018 interactions, consisting of 65.6% favourites, 31.9% boosts, and 2.5% replies.

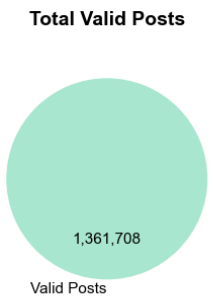
The `livefeeds.json` dataset contains Mastodon toots from all instances over approximately two weeks. The `boostersfavourites.json` and `replies.json` files provide data on boosts, favorites, and replies for these toots.

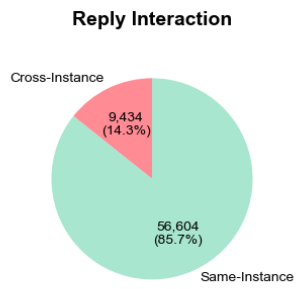
In Mastodon, a "toot" is a post made by a user. A "boost" is similar to a retweet on Twitter, allowing users to share a toot with their followers. A "favorite" is akin to a like, indicating appreciation for a toot. A "reply" is a response to another user's toot, facilitating conversations.

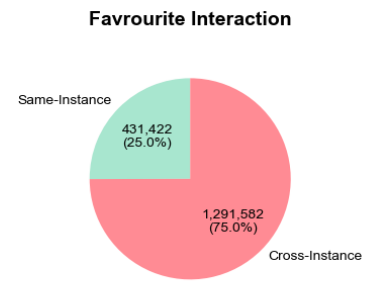
These interactions are essential for analyzing user engagement and content reach on Mastodon. By examining the `boostersfavourites.json` and `replies.json` files, researchers can gain insights into how content spreads and how users engage with each other on the platform.

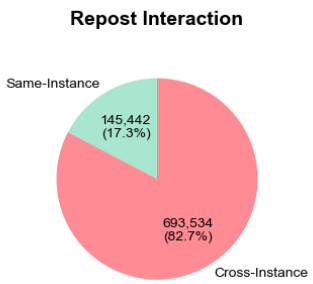
"Since each instance in the Fediverse maintains its own independent ID namespace, toots (posts) and user accounts from different instances may share identical local IDs. To ensure global uniqueness:

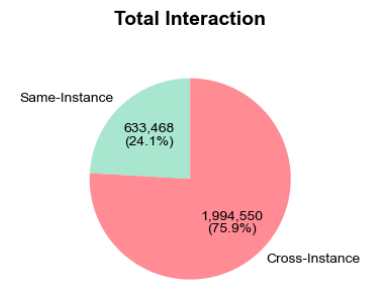
- **Toots** are uniquely identified using their `sid` (server-generated ID) combined with their instance-specific `url`
- **User Accounts** are uniquely identified through their canonical profile `url`, which inherently contains instance information."











Data-The Build of The Decentralized Dataset On Mastodon

Why to preprocess the dataset?

The current problems of the dataset

1. Noises in the dataset

Invalid Values （None/empty strings）

Duplicate Data

Inconsistent formats （non-uniform user-id formats,instance IDs with # suffixes）

2. Ambiguous analysis objects

Occassional errors in user behavior

Small-scale instances have no network influences
（Core influential instances : active users >=20）

3. Structure not aligned to the research

Mixed interactions between cross-instance and intra-instance

Lack of structural interaction relationship data

Data-The Build of The Decentralized Dataset On Mastodon

The process of data cleaning

1. Format Standardization

Instance IDs: Remove unnecessary suffixes (where present); unify formats

```
def extract_instance_id(url_or_sid):
    url_or_sid_str = str(url_or_sid).strip()
    if '#' in url_or_sid_str:
        url_or_sid_str = url_or_sid_str.split('#')[0]
    parsed = urlparse(url_or_sid_str)
    return parsed.netloc if parsed.netloc else ""
```

User IDs: Trim leading and trailing extra spaces

```
def extract_user_info(account):
    user_id = str(account.get("id", "")).strip()
    user_id = user_id if (user_id and user_id != "None") else ""
    return user_id
```

Timestamps: Convert all to YYYY-MM-DD HH:MM:SS format

2. Outlier Handling

Handling of None values: Resolve the error "NoneType has no attribute 'strip'"

```
def handle_none_value(value):
    return "" if value is None else value
```

Handling of invalid formats: Address abnormal data types in some fields (e.g., user information stored as lists instead of dictionaries, user IDs stored as numbers instead of strings)

```
def handle_invalid_type
  (field_data, field_type="user"):
    if not isinstance(field_data, (dict, str)):
        return ""
    if field_type == "user"
      and isinstance(field_data, dict):
        return
          str(field_data.get("id", "")).strip()
    return str(field_data).strip()
```

3. Duplicate Data Removal

Define a unique identifier rule (event_id)

Combine 4 core fields: interaction type, initiator user ID, recipient ID, precise timestamp

Generate unique identifiers

Remove duplicates based on the unique identifier

```
def remove_duplicates(interactions):
    unique_events = set()
    clean_interactions = []
    for interaction in interactions:
        event_id = generate_event_id(interaction)
        if event_id not in unique_events:
            unique_events.add(event_id)
            clean_interactions.append(interaction)
    return clean_interactions
```

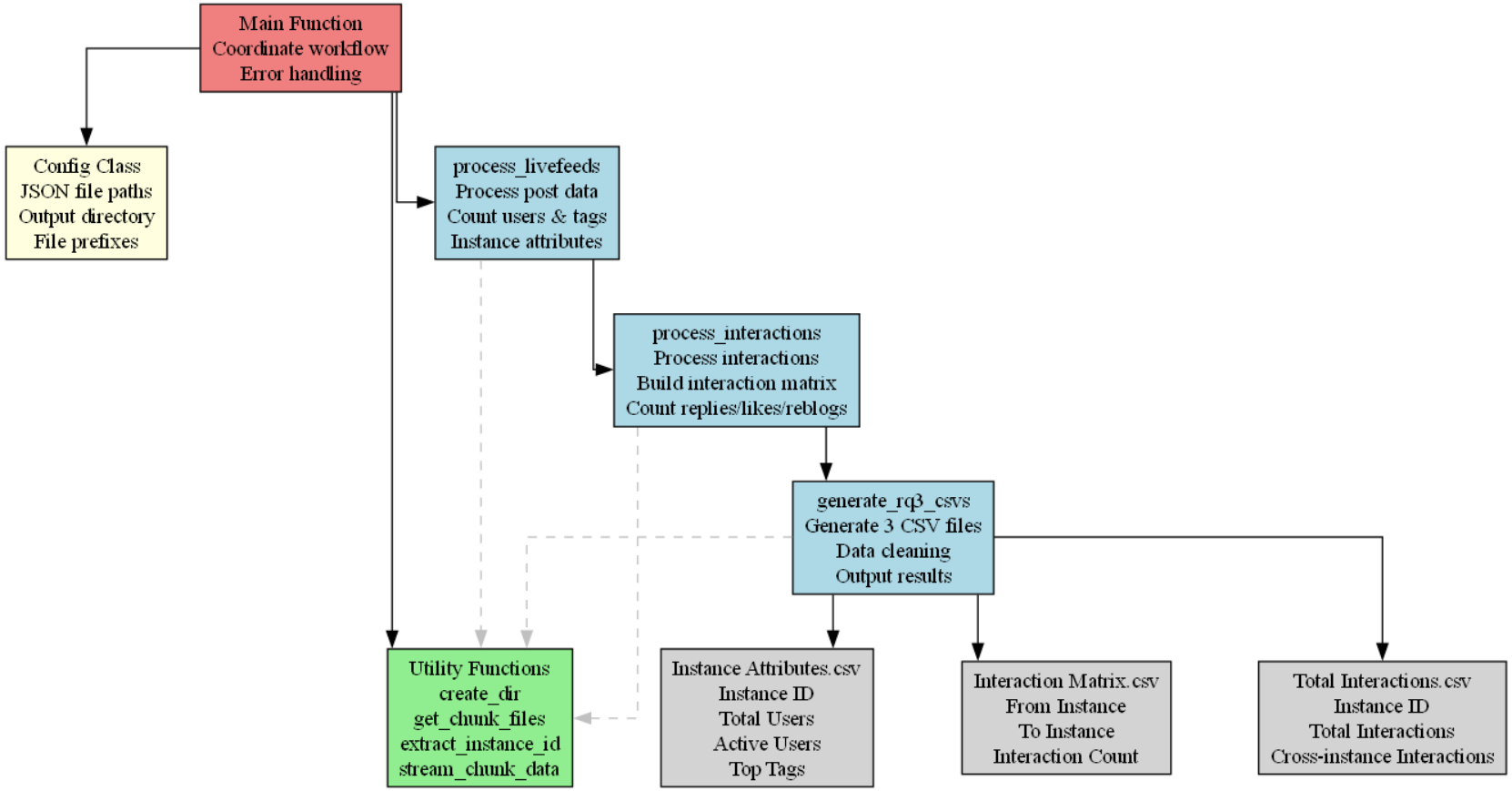
Data-The Build of The Decentralized Dataset On Mastodon

Extract key information from the existing dataset

1. Construction of the Basic Dataset

Consider interactions between all instances

Generate: interaction_table.csv, instance_attributes.csv , interaction_matrix.csv, and instance_interaction_stats.csv



Data-The Build of The Decentralized Dataset On Mastodon

Output: Standardized Dataset

1. interaction_table.csv

Field Name	Meaning	Type
event_id	Unique Event ID	string / int
timestamp	UTC Timestamp	datetime
from_user_id	Initiator User ID	string
from_instance	Instance Domain of Initiator	string
to_user_id	Recipient User ID	string
to_instance	Instance Domain of Recipient	string
interaction_type	Interaction Type (Boost, Favourite, Bookmark)	enum
is_cross	Cross-Instance or Not	bool
weight (optional)	Interaction Intensity	int / float

Statistic Item	Value
Number of Nodes (Instances)	8,020
Number of Edges (Instance Interaction Pairs)	101,218
Number of Users (Participating in Interactions)	574,049
Total Interaction Records	2,628,018

```
event_id,timestamp,from_user_id,from_instance,to_user_id,to_instance,interaction_type,is_cross,weight
reply_anon_id_57_anon_id_57_2024-12-04-00-00-00_f072cbe0,2024-12-04T00:00:00Z,anon_id_5701077538f5d117f3,mastodon.social,anon_id_5701077538f5d117f3,mastodon.social,reply,False,1
reply_anon_id_2f_anon_id_2f_2024-12-04-00-00-00_b3ed5c99,2024-12-04T00:00:00Z,anon_id_2ff311fb125c9fe331,mastodon.social,anon_id_2ff311fb125c9fe331,mastodon.social,reply,False,1
reply_anon_id_7b_anon_id_7b_2024-12-04-00-00-00_c0d9efa1,2024-12-04T00:00:00Z,anon_id_7b1ebf8d574c15a128,mastodon.social,anon_id_7b1ebf8d574c15a128,mastodon.social,reply,False,1
reply_anon_id_a1_anon_id_a1_2024-12-04-00-00-00_bf81e2e8,2024-12-04T00:00:00Z,anon_id_a134a14a27b9933674,mastodon.social,anon_id_a134a14a27b9933674,mastodon.social,reply,False,1
reply_anon_id_be_anon_id_be_2024-12-04-00-00-00_456fc230,2024-12-04T00:00:00Z,anon_id_be2854dcc5241cb3f5,mastodon.social,anon_id_be2854dcc5241cb3f5,mastodon.social,reply,False,1
reply_anon_id_fa_anon_id_fa_2024-12-04-00-00-00_a2ed9a96,2024-12-04T00:00:00Z,anon_id_fa9ed44855b96f796b,mastodon.social,anon_id_fa9ed44855b96f796b,mastodon.social,reply,False,1
reply_anon_id_7b_anon_id_7b_2024-12-04-00-00-00_23fffc24,2024-12-04T00:00:00Z,anon_id_7b1ebf8d574c15a128,mastodon.social,anon_id_7b1ebf8d574c15a128,mastodon.social,reply,False,1
reply_anon_id_7b_anon_id_7b_2024-12-04-00-00-00_11c2dd2b,2024-12-04T00:00:00Z,anon_id_7b1ebf8d574c15a128,mastodon.social,anon_id_7b1ebf8d574c15a128,mastodon.social,reply,False,1
```

Data-The Build of The Decentralized Dataset On Mastodon

Output: Standardized Dataset

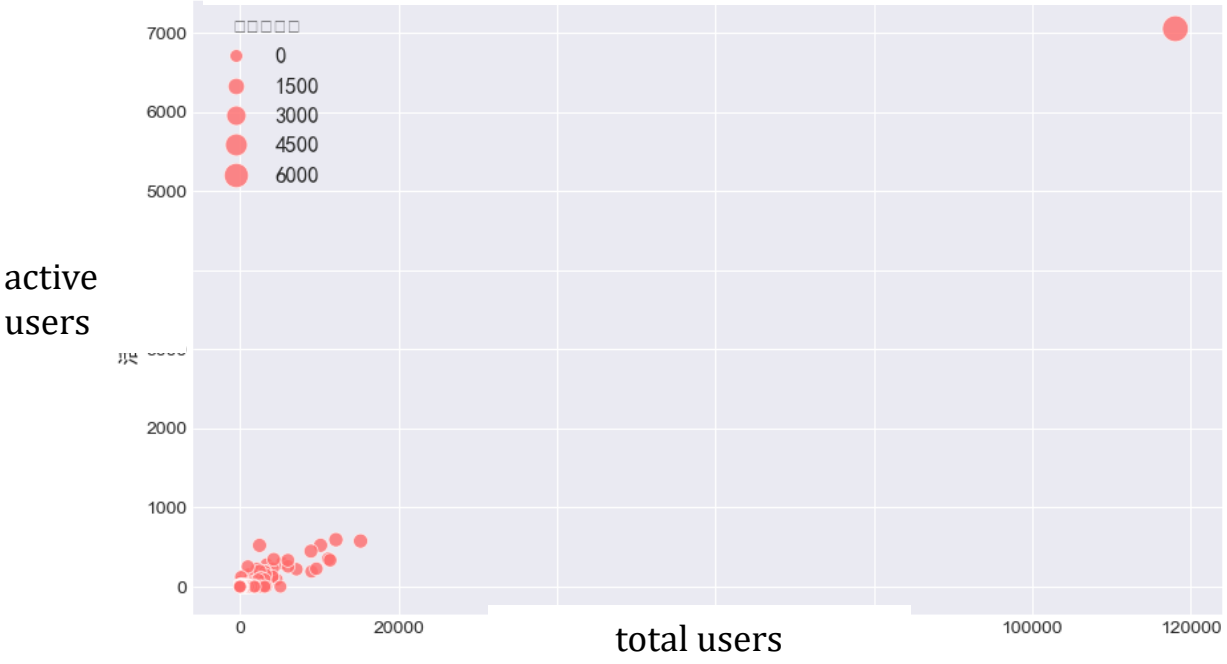
2. instance_attributes.csv

Field Name	Explanation
instance_id	Unique identifier of the instance
users_number	Users who have posted or interacted in the instance
active_users_number	All active users in the instance
tag	Top 5 most frequently discussed topics in the instance

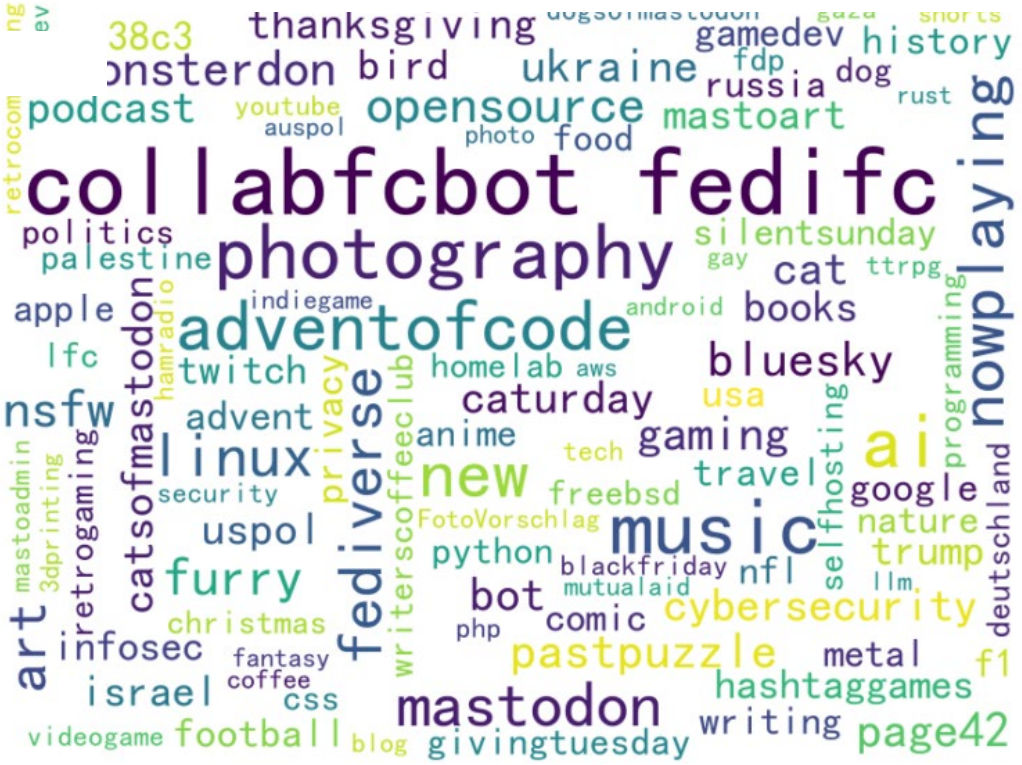
tags:extracted from----


```
"tags": [
    {
        "name": "crowcontent",
        "url": "https://chaos.social/tags/crowcontent"
    },
    {
        "name": "crows",
        "url": "https://chaos.social/tags/crows"
    },
    {
        "name": "corvids",
        "url": "https://chaos.social/tags/corvids"
    },
    {
        "name": "photography",
        "url": "https://chaos.social/tags/photography"
    }
],
```


Correlation between Total Number of Users and Active Users



Word Cloud of Popular Hashtags in the Instance



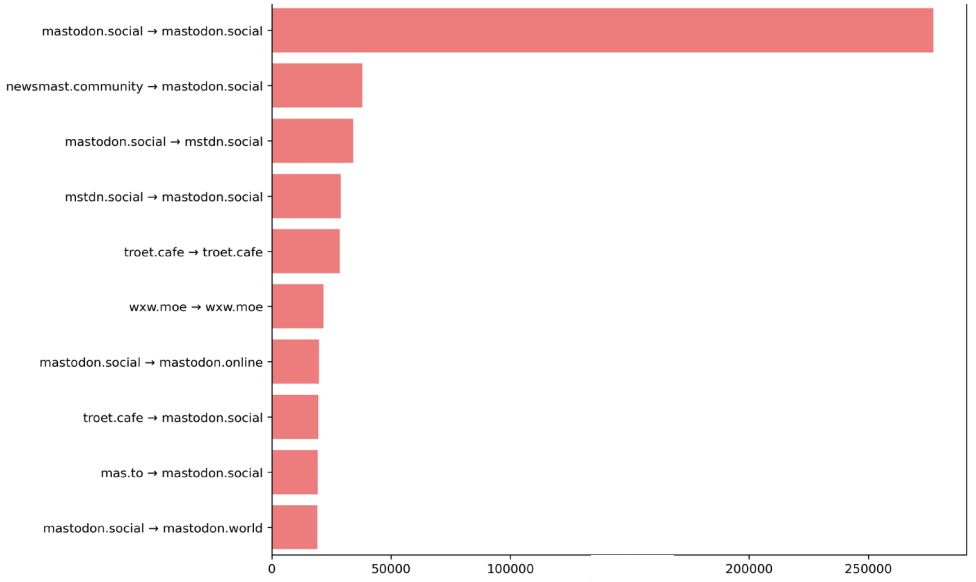
Data-The Build of The Decentralized Dataset On Mastodon

Output: Standardized Dataset

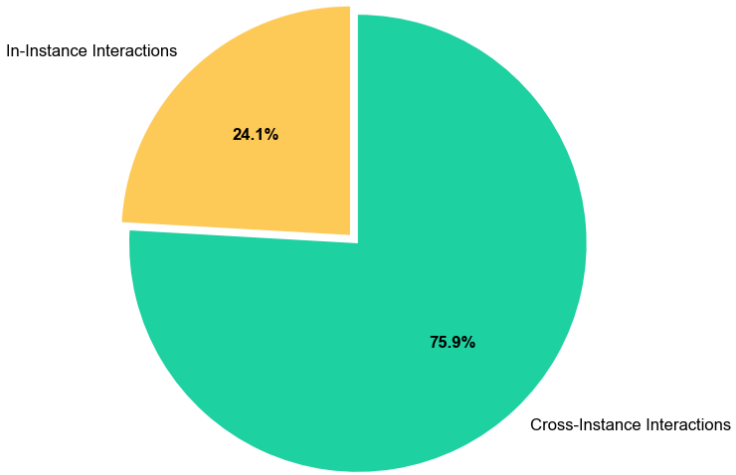
3. interaction_matrix.csv

Field Name	Meaning	Explanation
from_instance	Source Instance	the instance initiaing the interaction
to_instance	Target Instance	the instance receiving the interaction
reply_count	Reply Count	user replies to a post
favrourite_count	Favrourite Count	user favrourites a post
reblog_count	Reblog Count	user reblogs a post
interaction_count	Total Interaction Count	reply+favrourite+reblog

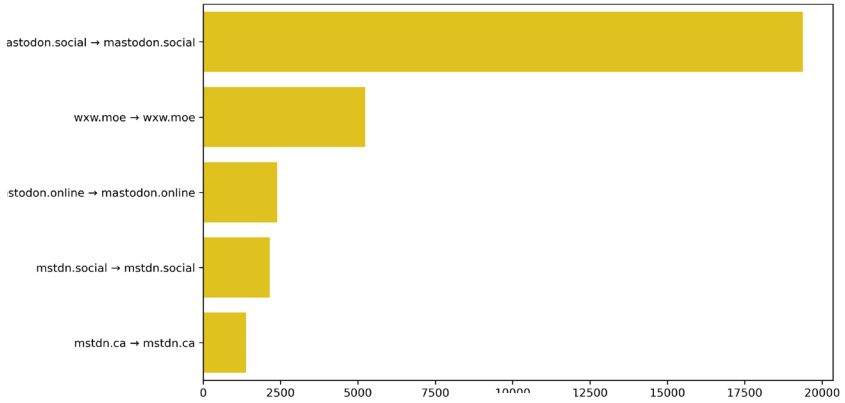
Total Interaction Count of the Top 10 Instances



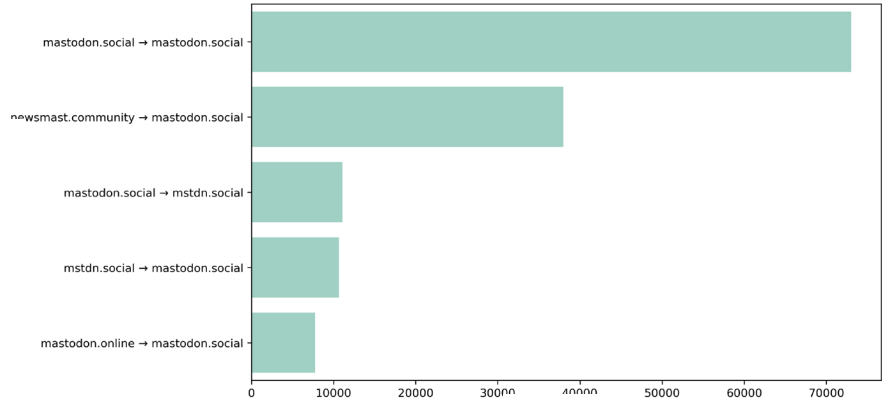
Proportion of Cross-Instance vs In-Instance in Total Interactions



Top 5 Instances Reply Count



Top 5 Instances Boost Count



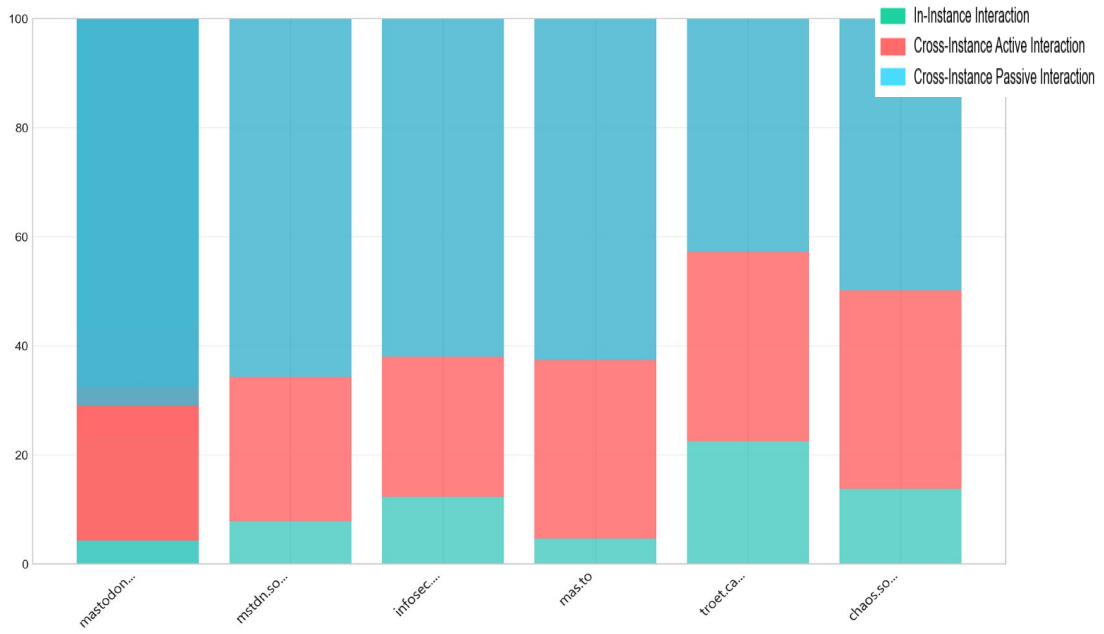
Data-The Build of The Decentralized Dataset On Mastodon

Output: Standardized Dataset

4. instance_interaction_stat.csv

Field Name	Meaning
instance_id	Instance ID
in_reply_count	Internal Reply Count
in_boost_count	Internal Boost Count
in_fav_count	Internal Favorite Count
in_interaction_count_total	Total Internal Interactions
ex_active_reply_count	Cross-instance Active Reply Count
ex_active_boost_count	Cross-instance Active Boost Count
ex_active_fav_count	Cross-instance Active Favorite Count
ex_active_interaction_count_total	Total Cross-instance Active Interactions
ex_passive_reply_count	Cross-instance Passive Reply Count
ex_passive_boost_count	Cross-instance Passive Boost Count
ex_passive_fav_count	Cross-instance Passive Favorite Count
ex_passive_interaction_count_total	Total Cross-instance Passive Interactions
ex_interaction_count_total	Total Cross-instance Interactions

Distribution of Instance Interaction Types (by Percentage)



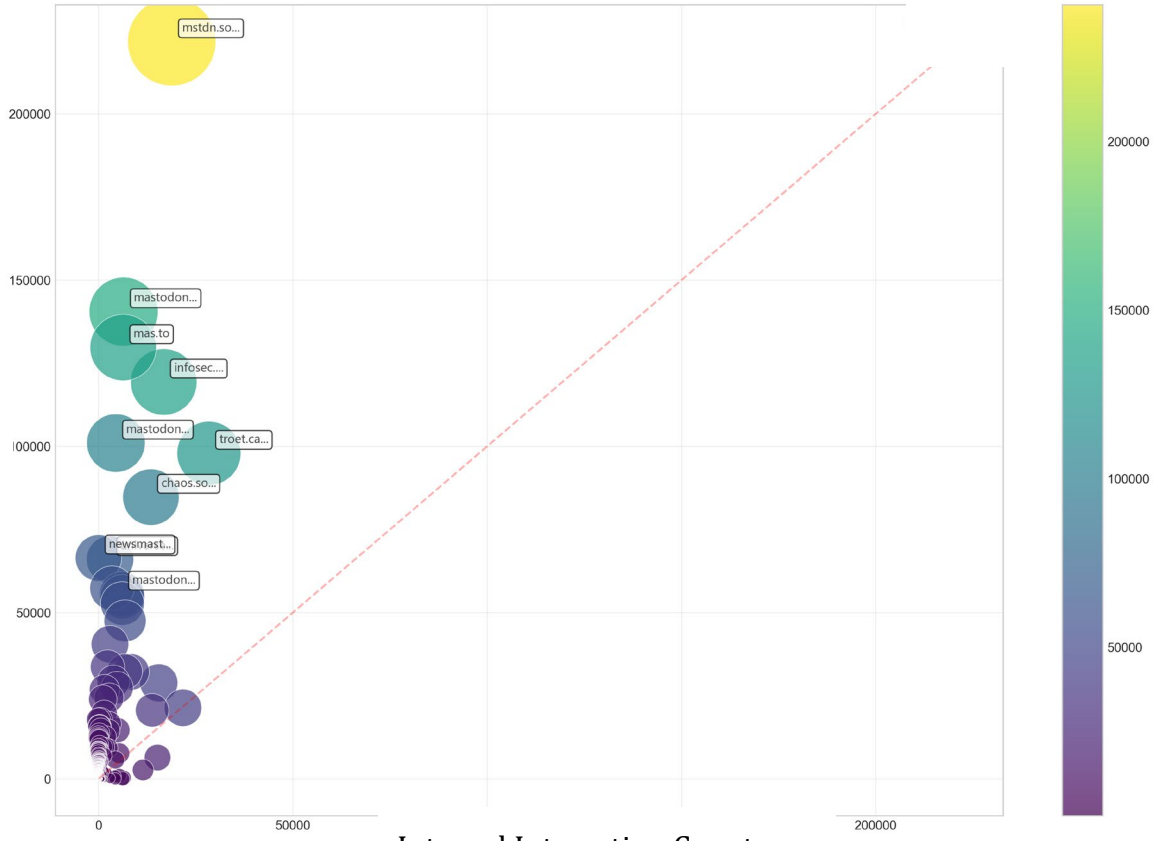
Data-The Build of The Decentralized Dataset On Mastodon

Output: Standardized Dataset

4. instance_interaction_stat.csv

Field Name	Meaning
instance_id	Instance ID
in_reply_count	Internal Reply Count
in_boost_count	Internal Boost Count
in_fav_count	Internal Favorite Count
in_interaction_count_total	Total Internal Interactions
ex_active_reply_count	Cross-instance Active Reply Count
ex_active_boost_count	Cross-instance Active Boost Count
ex_active_fav_count	Cross-instance Active Favorite Count
ex_active_interaction_count_total	Total Cross-instance Active Interactions
ex_passive_reply_count	Cross-instance Passive Reply Count
ex_passive_boost_count	Cross-instance Passive Boost Count
ex_passive_fav_count	Cross-instance Passive Favorite Count
ex_passive_interaction_count_total	Total Cross-instance Passive Interactions
ex_interaction_count_total	Total Cross-instance Interactions

Analysis of Instance Interaction Patterns (Excluding the Largest Instance)



Internal Interaction Count

Data-The Build of The Decentralized Dataset On Mastodon

Core Instance Filtering

5. Filter by Active User Criteria

Active User Definition:
Number of posts >= 1 && Number of interactions >= 3

Only consider interactions between active users

6. Filter by Core Instance Criteria

Core Instance Definition:
Number of active users >= 20

Only filter interactions between active instances

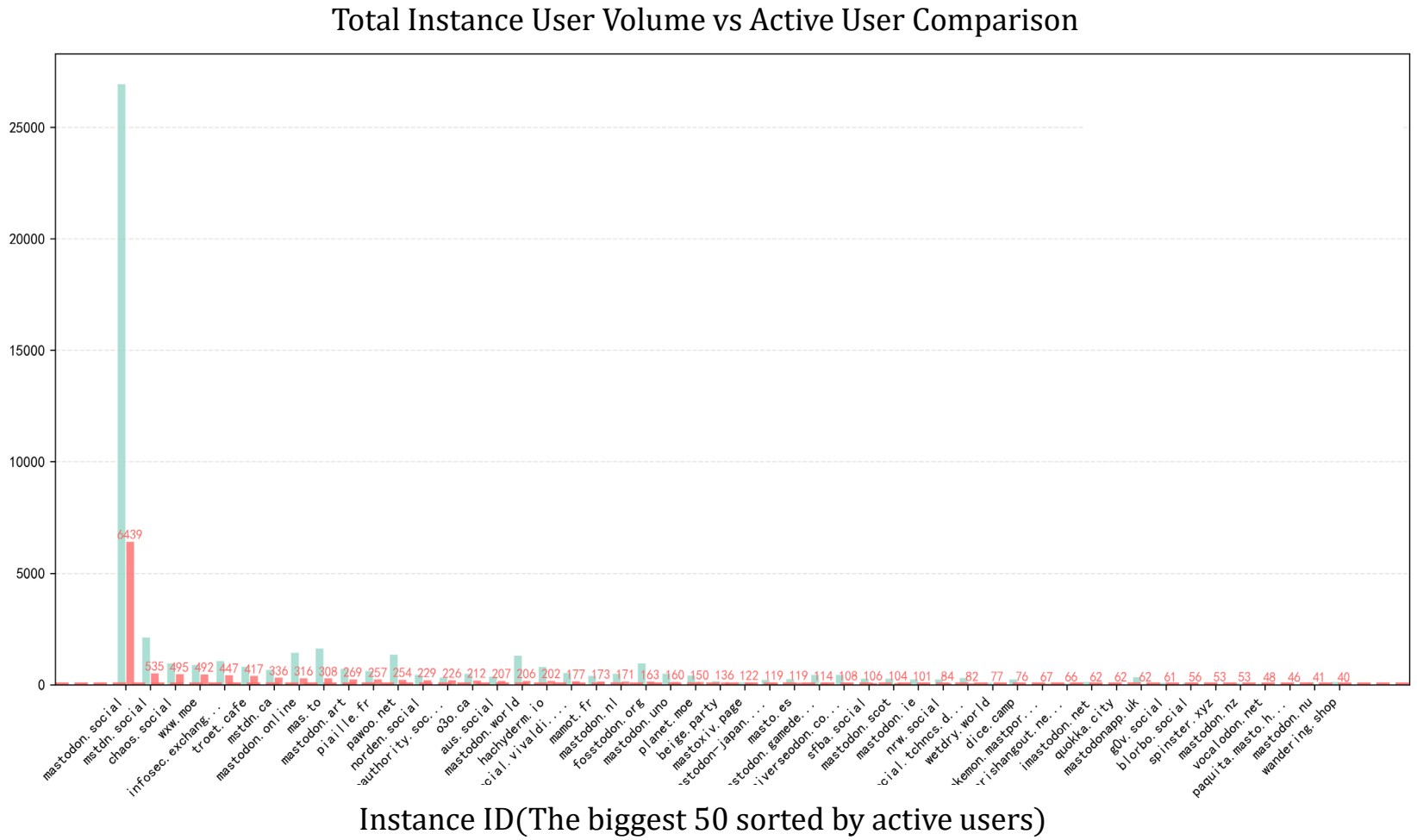


Figure: Long-tail distribution of total users and active users across different instances

RQ1 Behavioral Motivation Layer: How does instance size affect cross-instance interaction?

Data Crawling and Processing

FOR RQ1:

Sort by total number of users

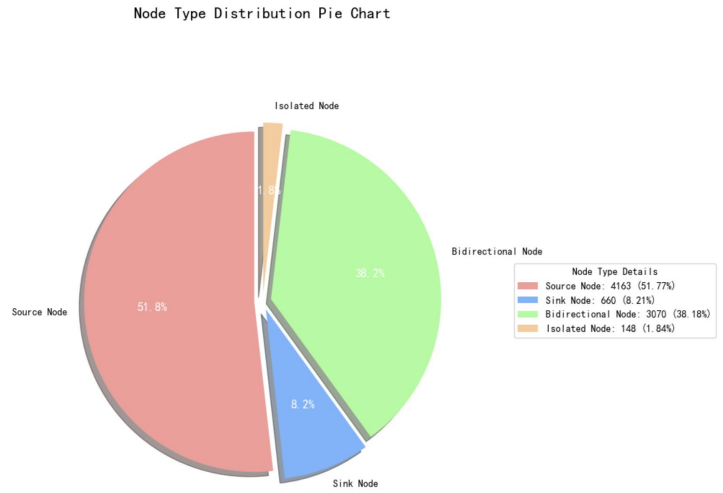
Sorted by active users

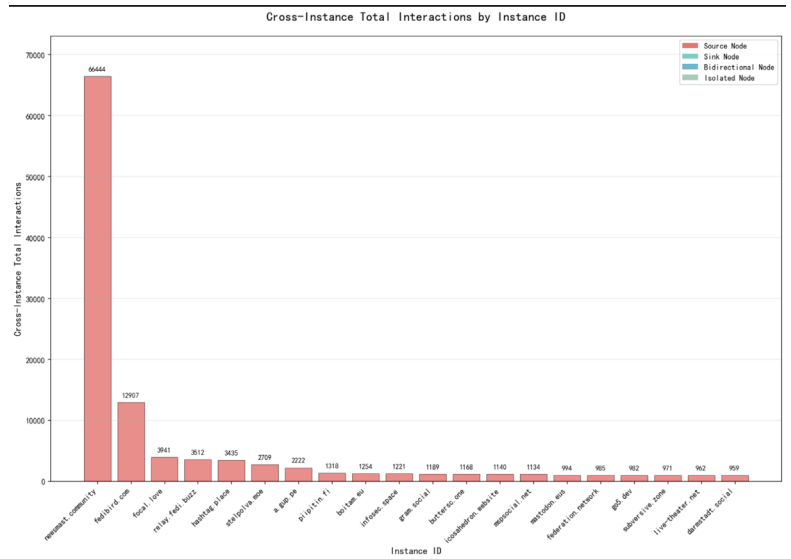
Sorted by internal interactions

Sorted by cross-instance interaction count

Sorted by cross-instance interaction ratio

Further analysis of instance interactions identifies source nodes, sink nodes, isolated nodes, and bidirectional nodes.





RQ1 Behavioral Motivation Layer: How does instance size affect cross-instance interaction?

Preliminary Analysis of Instance Size

FOR RQ1:

A total of 8,963 instances appeared in our dataset.

The top 10 instances account for 35.4% of all users.

Users ≥ 10000 : 6

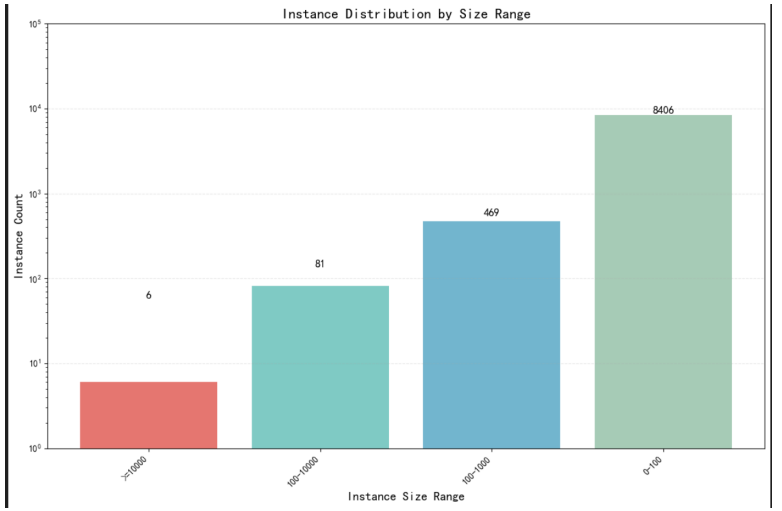
Users ≥ 1000 : 81

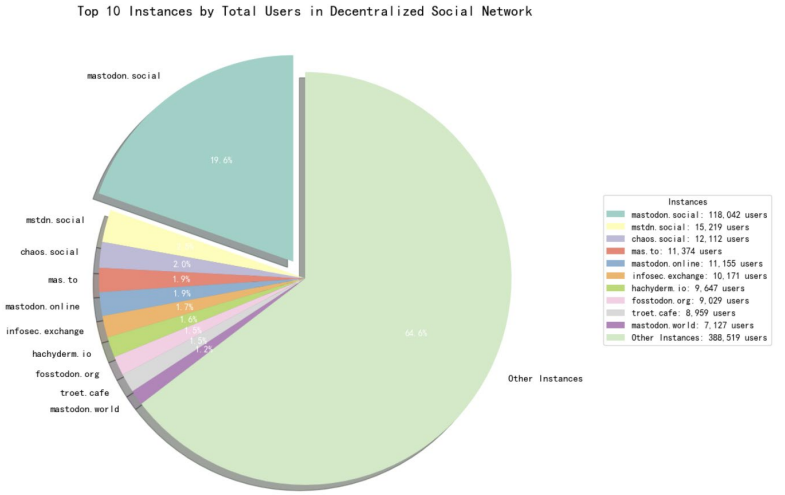
Users ≥ 100 : 469

Users ≤ 100 : 8406

Gini: 0.928

The distribution of instance sizes is extremely uneven.





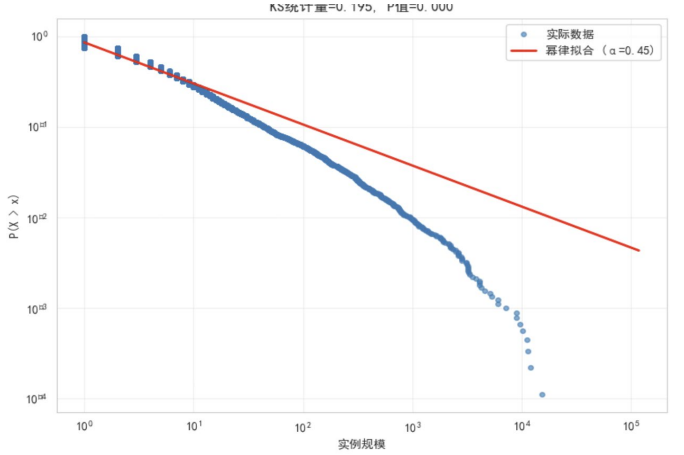
RQ1 Behavioral Motivation Layer: How does instance size affect cross-instance interaction?

Fitting Analysis

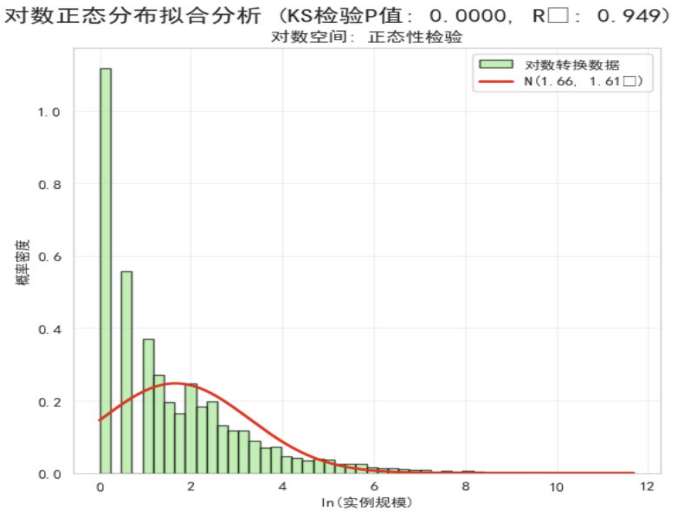
	R^2	P-Value
Power-law distribution	0.968	**<0.0001**
log-normal distribution	0.949	**<0.0001**
exponential distribution	-0.845	**<0.0001**

The data does not conform to a strict power-law distribution, but exhibits characteristics of a power-law distribution.

Power-law distribution：



log-normal distribution



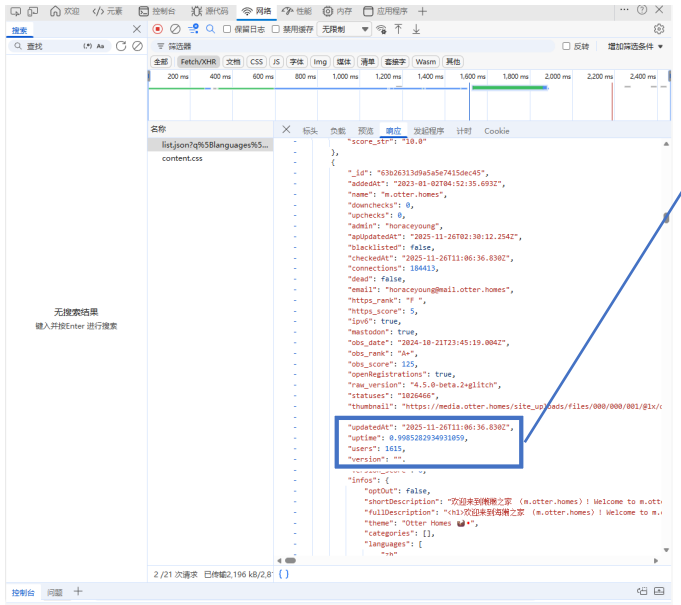
RQ1 Behavioral Motivation Layer: How does instance size affect cross-instance interaction?

Data Crawling and Processing

To more accurately reflect the actual user scale distribution of instances, we crawl the corresponding instance scales from instance.social based on the instance list and perform the fitting process again as described above.

```
"uptime": 0.9985282934931059,
"users": 1615,
"version": "",
```

```python
def fetch_instance_user_count(instance_domain: str) -> Dict:
    """
    获取单个实例的用户数量信息

    参数:
        instance_domain (str): 实例域名

    返回:
        Dict: 实例信息字典
    """
    # API地址 - 使用instances.social的API
    api_url = "https://instances.social/api/1.0/instances/list"

    # 请求参数 - 搜索特定实例
    params = {
        'q[name]': instance_domain,
        'count': 1  # 只获取最相关的结果
    }

    # 请求头
    headers = {
        'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.124 Safari/537.36',
        'Accept': 'application/json',
    }
```



target_d	matched_users	up	open_regi	languages	match_status
mastodon.	mastodon.	2972605	TRUE	TRUE	en	完全匹配
bots.fyri	bots.fyri	424812	TRUE	TRUE		完全匹配
mastodon.	mastodon.	277814	TRUE	TRUE		完全匹配
mstdn.soc	mstdn.soc	267432	TRUE	TRUE	ab, aa, a	完全匹配
pravda.me	pravda.me	228960	TRUE	TRUE		完全匹配
mastodon.	mastodon.	193552	TRUE	TRUE	nl, en	完全匹配
techhub.s	techhub.s	84901	TRUE	TRUE	en	完全匹配
universec	universec	81712	TRUE	TRUE	en	完全匹配
mastodona	mastodona	80160	TRUE	TRUE	ab, aa, a	完全匹配
pixelfed.	pixelfed.	79123	TRUE	TRUE		完全匹配
infosec.e	infosec.e	78767	TRUE	TRUE		完全匹配
mastodon.	mastodon.	78251	TRUE	TRUE	it	完全匹配
mastodon.	mastodon.	73202	TRUE	TRUE		完全匹配
social.vi	social.vi	72316	TRUE	TRUE		完全匹配
c.im	c.im	68002	TRUE	TRUE	en	完全匹配
gram.soci	gram.soci	61239	TRUE	TRUE		完全匹配
pixtagram	gram.soci	61239	TRUE	TRUE		部分匹配
hachyderm	hachyderm	57039	TRUE	TRUE		完全匹配
brighteor	brighteor	55863	TRUE	TRUE		完全匹配
dev.brigh	dev.brigh	54476	TRUE	TRUE		完全匹配
mstdn.par	mstdn.par	53125	TRUE	TRUE	en	完全匹配
aethy.com	aethy.com	49149	TRUE	TRUE	en	完全匹配
m.cmx.im	m.cmx.im	48202	TRUE	TRUE	zh	完全匹配
troet.caf	troet.caf	47751	TRUE	TRUE	de	完全匹配

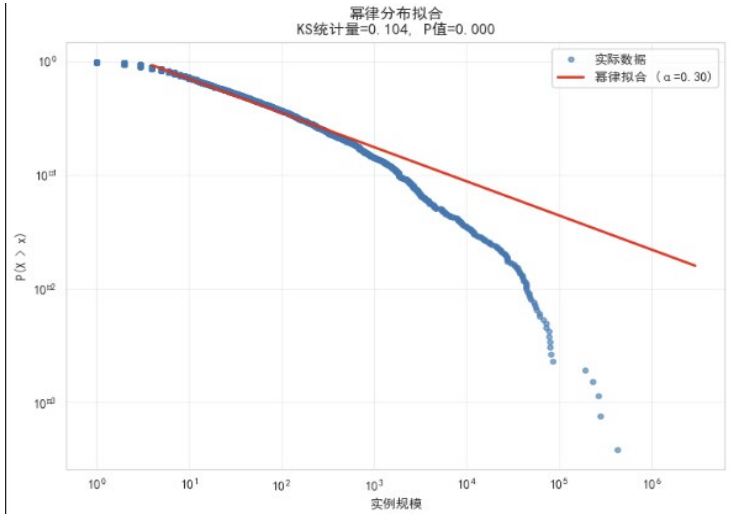
RQ1 Behavioral Motivation Layer: How does instance size affect cross-instance interaction?

Fitting Analysis

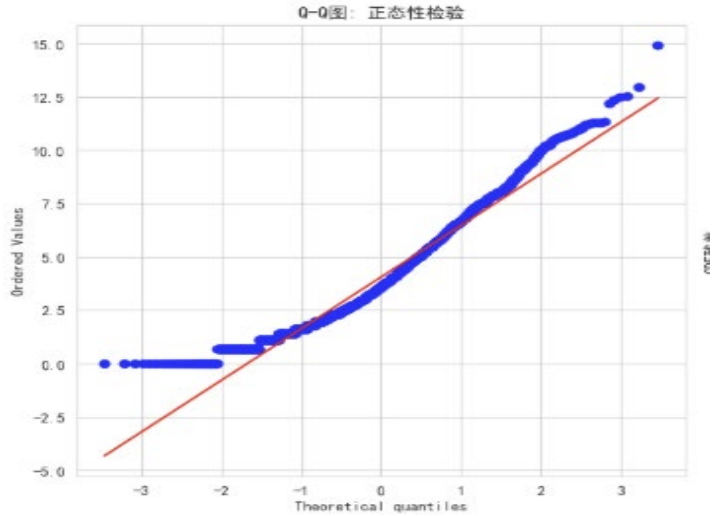
	R^2	P-Value
Power-law distribution	0.989	**<0.0001**
log-normal distribution	0.977	**<0.0001**
exponential distribution	-1.275	**<0.0001**

The data does not conform to a strict power-law distribution, but exhibits characteristics of a power-law distribution.

Power-law distribution：



log-normal distribution



RQ2 Semantic Structure Layer: How Languages and Topics Jointly Shape the Semantic Community Structure

Community Testing

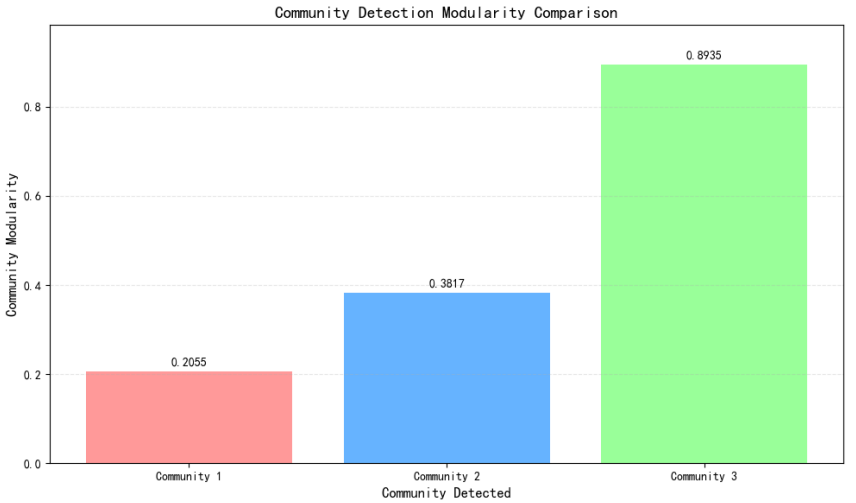
Using instances as nodes

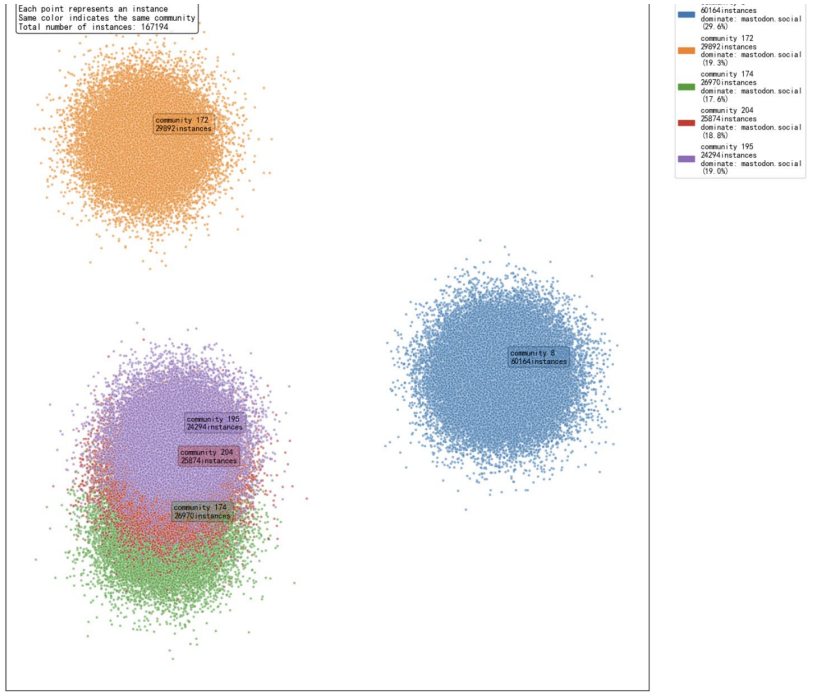
Ignore internal interactions： 0.2055

Consider internal interactions： 0.3817

Users as nodes： 0.8935

This implies that in decentralized social networks, the true social boundaries lie at the user level, not at the instance level.





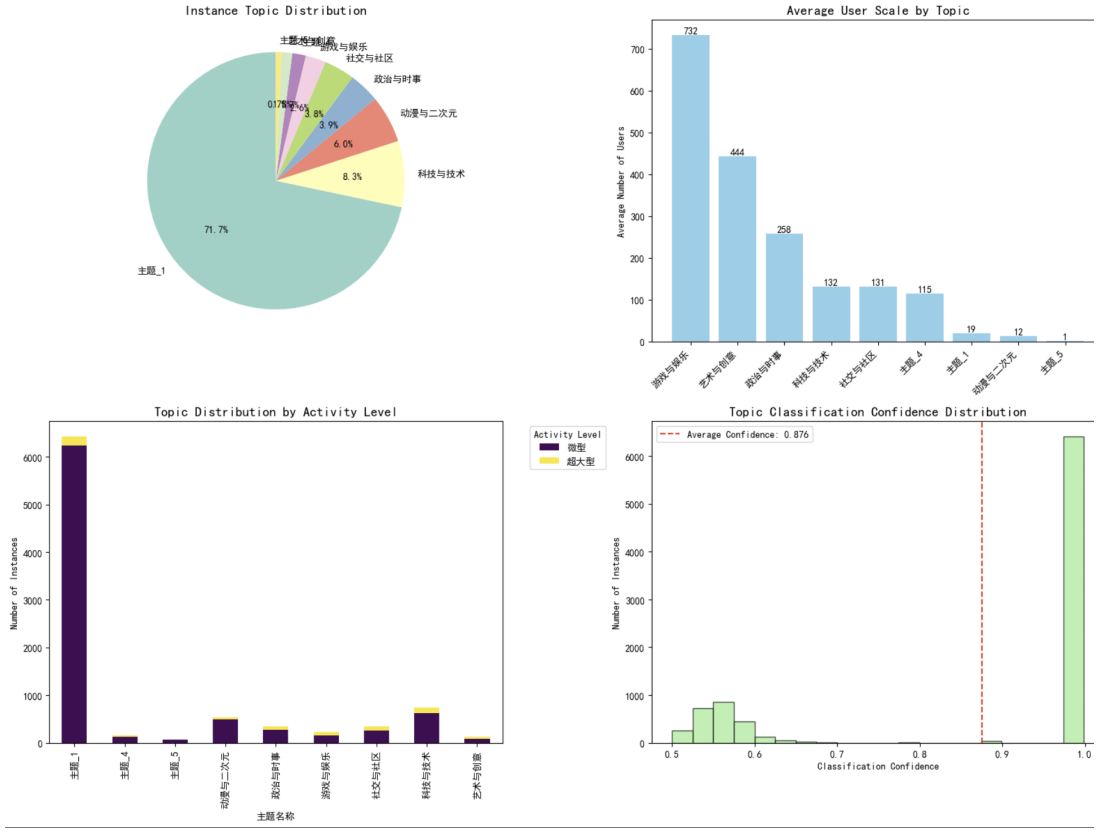
RQ2 Semantic Structure Layer: How Languages and Topics Jointly Shape the Semantic Community Structure

Data Crawling and Processing

FOR RQ2:

Use natural language processing models to derive instance themes.

Dataset:“instances_sorted_by_activeusers”



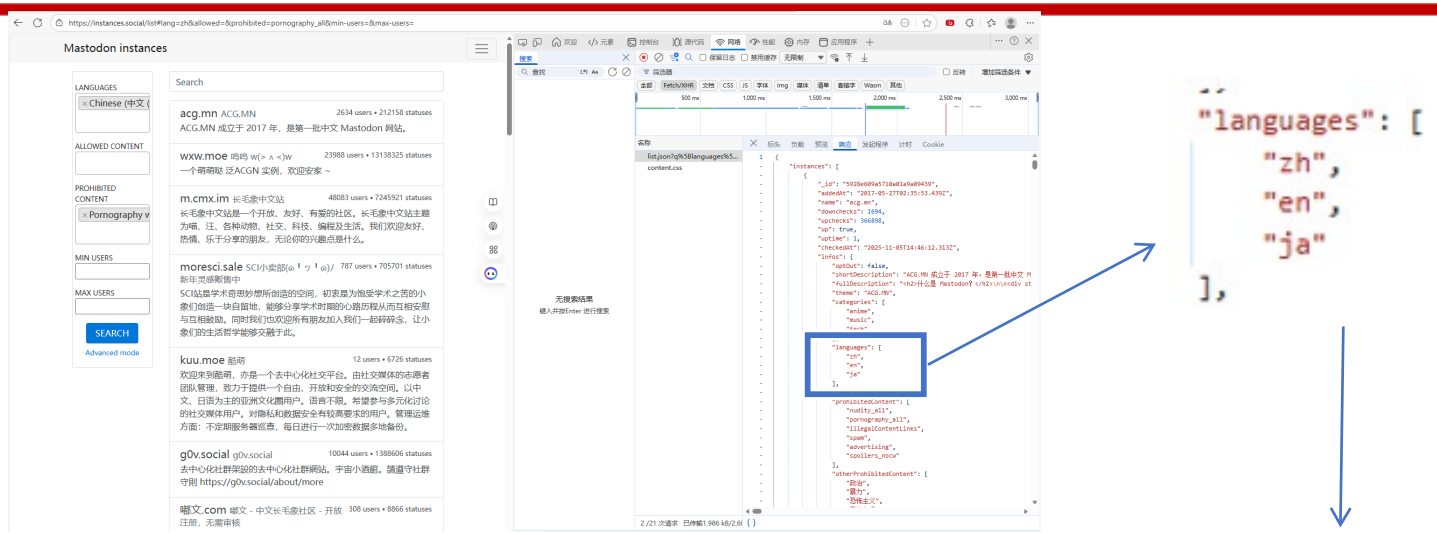
Processing the subject tags for all instances

Application of K-means Clustering

Analyze statistics and generate charts

RQ2 Semantic Structure Layer: How Languages and Topics Jointly Shape the Semantic Community Structure

Data Crawling and Processing



FOR RQ2:

The language of the instance

1） For instances containing language data within instance.social, directly scrape the language information using a web crawler.

```python
def scrape_instances_with_detected_language() -> List[Dict]:
    """
    从 instances.social API 爬取实例信息，并通过描述文本检测语言
    """
    api_url = "https://instances.social/list.json"

    headers = {
        'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36',
        'Accept': 'application/json',
    }

    all_instances = []

    print("开始爬取 instances.social 数据...")

    try:
        response = requests.get(api_url, headers=headers, timeout=60)

        if response.status_code != 200:
            print(f"请求失败，状态码: {response.status_code}")
            return []

        data = response.json()
        instances = data.get('instances', [])
        print(f"成功获取 {len(instances)} 个实例")
```

RQ2 Semantic Structure Layer: How Languages and Topics Jointly Shape the Semantic Community Structure

Data Crawling and Processing

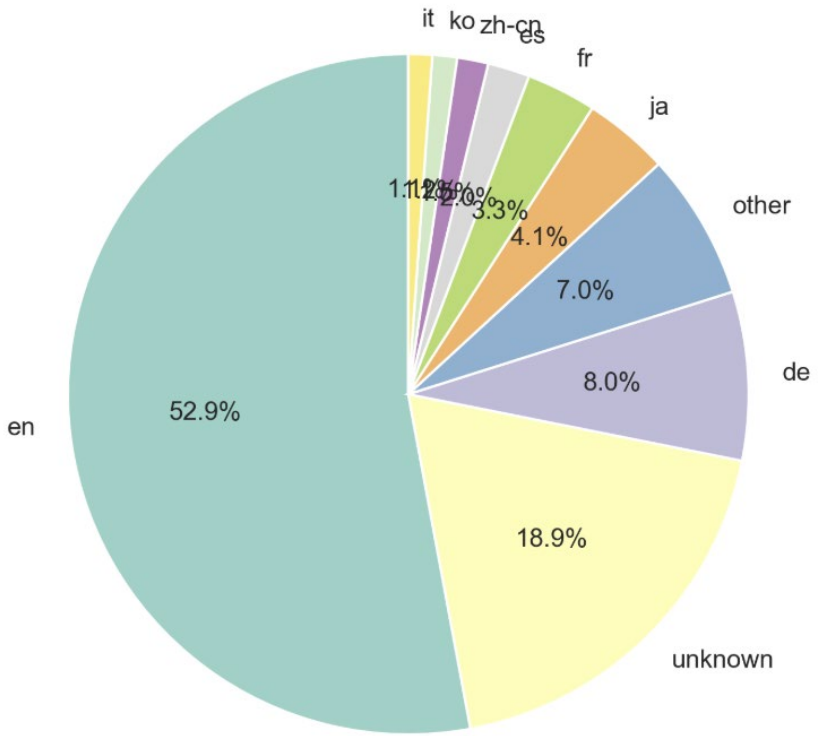
FOR RQ2:

The language of the instance

2） For instances in instance.social that do not specify a language, determine their language based on the instance description.

```
# 获取描述文本用于语言检测
short_desc = instance.get('infos', {}).get('shortDescription', '')
full_desc = instance.get('infos', {}).get('fullDescription', '')

# 合并描述文本
description_text = f"{short_desc} {full_desc}"
cleaned_text = clean_text(description_text)

# 检测语言
detected_language = "unknown"
language_source = "no_text"

if len(cleaned_text) > 20:  # 只有文本足够长时才检测
    detected_language = detect_language_from_text(cleaned_text)
    language_source = "detected"
    detected_count += 1
elif declared_languages:
    detected_language = primary_declared_language
    language_source = "declared"
else:
    no_language_count += 1
    language_source = "unknown"

instance_info = {
    'name': instance.get('name'),
    # 官方声明的语言
    'declared_languages': ', '.join(declared_languages),
    'primary_declared_language': primary_declared_language,
    'declared_language_count': len(declared_languages),
    # 检测到的语言
    'detected_language': detected_language,
    'language_source': language_source,
    # 描述文本信息（用于验证）
    'short_description': clean_text(short_desc)[:100],  # 只保留前100字符用于验证
    'text_length': len(cleaned_text)
}
all_instances.append(instance_info)
processed_count += 1

if processed_count % 100 == 0:
    print(f"已处理 {processed_count} 个实例...")
```

3） Compile the data and conduct a preliminary analysis.



RQ1 Behavioral Motivation Layer: How does instance size affect cross-instance interaction?

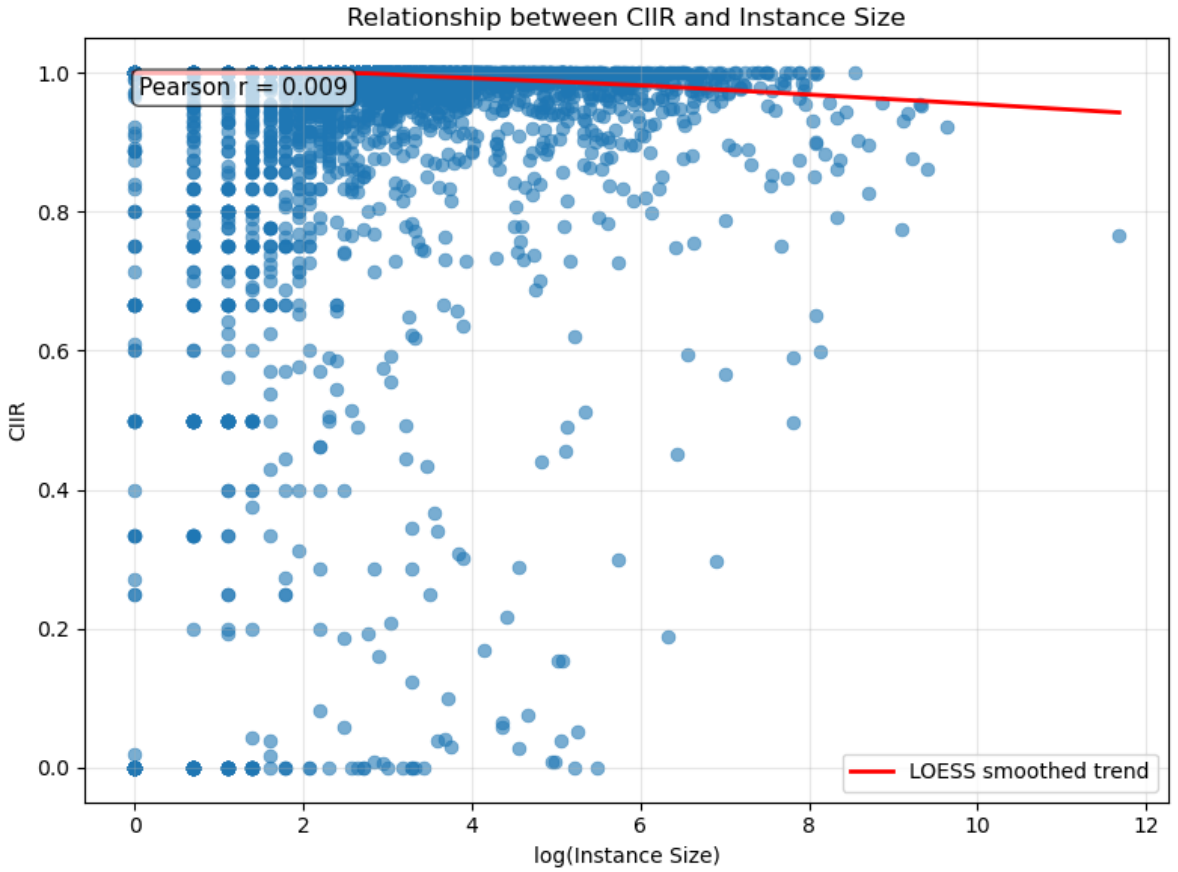
Key Definitions and Preliminary Analysis

Key Definitions

- Instance size：number of registered users
- Key metric：Cross-Instance Interaction Rate (CIIR)

Preliminary Analysis

- Scatter plot with x = log(instance size) and y = CIIR to visualize the relationship between instance size and CIIR
- Add a Loess curve to reveal potential nonlinear trends
- Pearson correlation test to assess linear dependence



Almost no linear correlation

RQ1 Behavioral Motivation Layer: How does instance size affect cross-instance interaction?

Group-based CIIR Comparison

Grouping

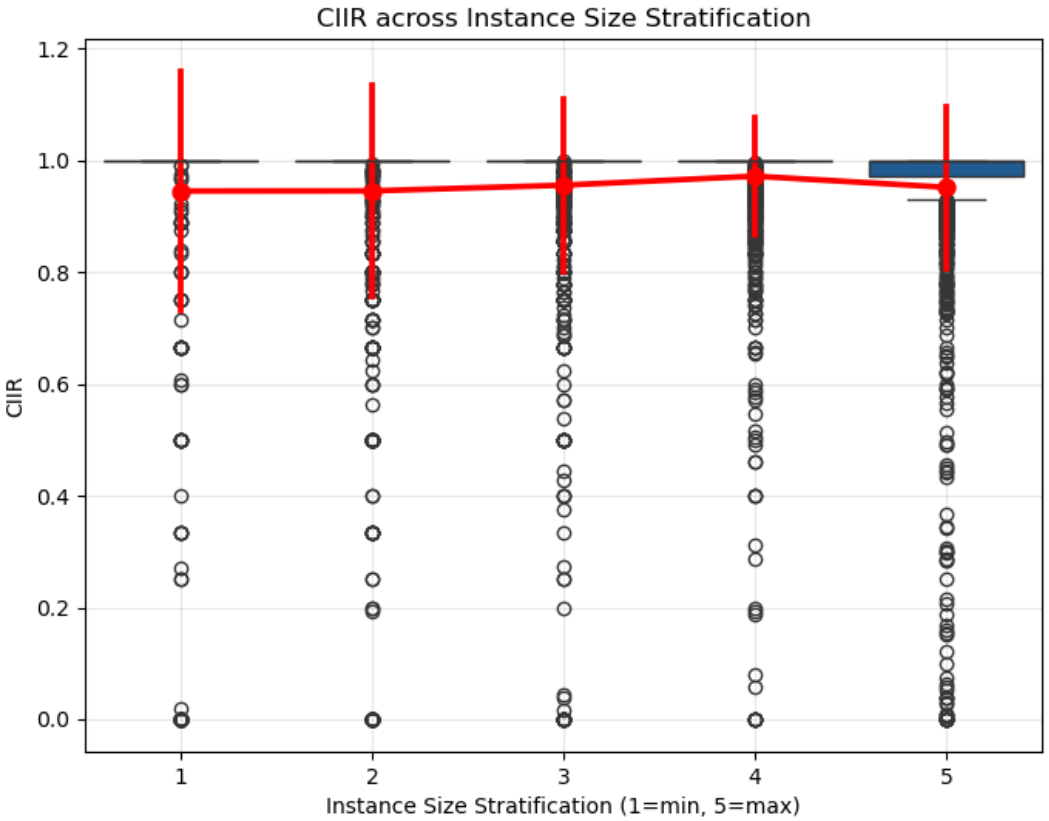
- Group instances by size (Level 1–5) to prevent extreme ratios in small instances

Boxplots

- Visualize CIIR with boxplots

Test

- Use ANOVA to test group differences
- Apply Tukey's HSD test for pairwise comparisons



ANOVA

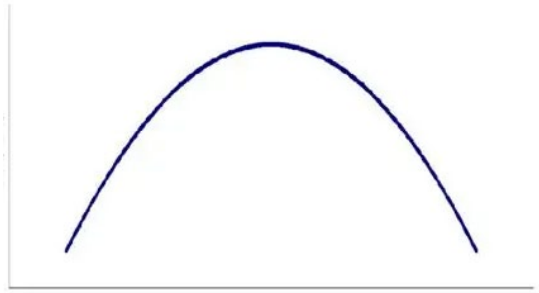
$F=6.88,\ p<0.001$

RQ1 Behavioral Motivation Layer: How does instance size affect cross-instance interaction?

Group-based CIIR Comparison

Tukey HSD test Results

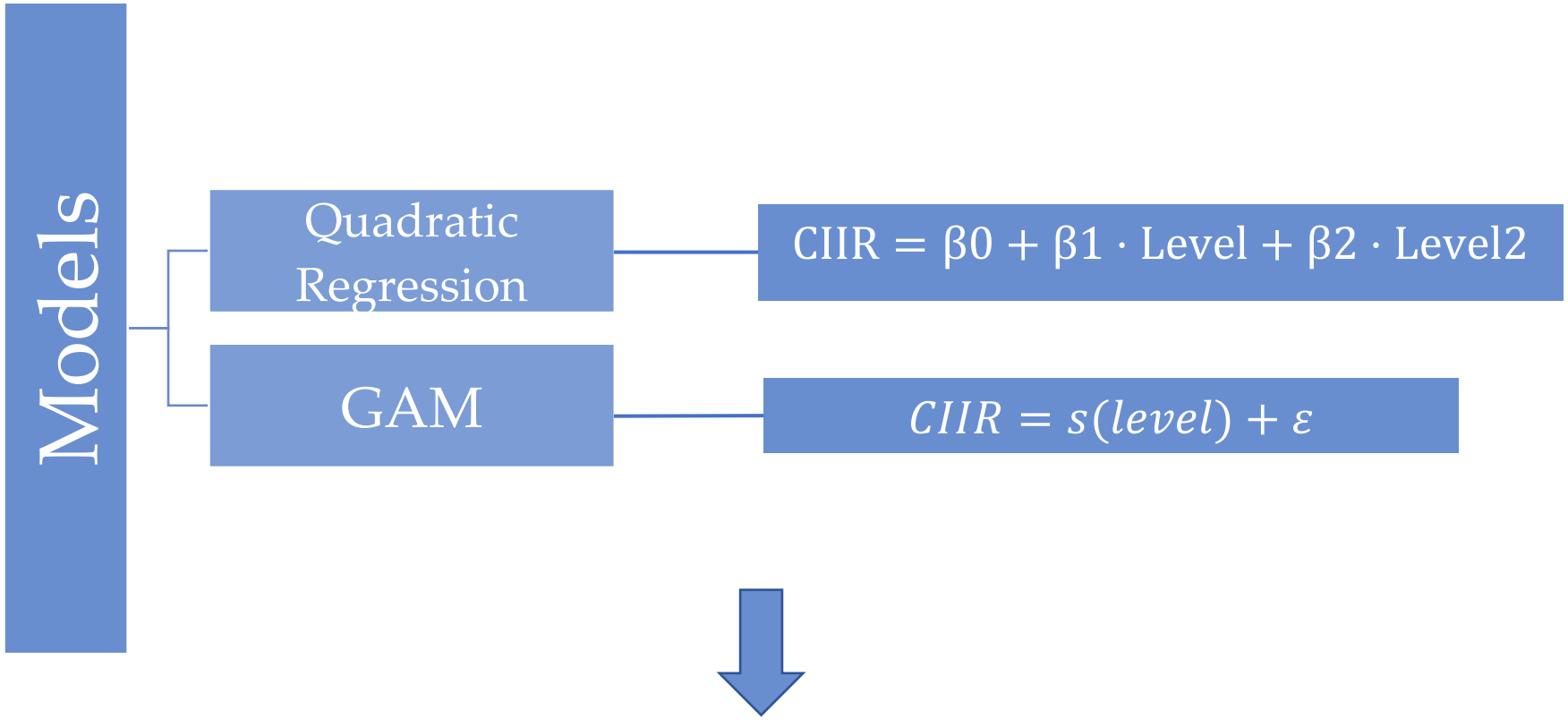
Comparison	Mean Difference	Adjusted p-value	Significance	Interpretation
1 vs 4	+0.0266	0.0001	✅ Significant	CIIR in Level 4 is significantly higher than Level 1
2 vs 4	+0.0264	<0.0001	✅ Significant	CIIR in Level 4 is higher than Level 2
4 vs 5	-0.0199	0.0051	✅ Significant	CIIR in Level 5 is slightly lower than Level 4
Other pairs	–	$p > 0.05$	❌ Not significant	No statistical difference



Inverted U-shaped

RQ1 Behavioral Motivation Layer: How does instance size affect cross-instance interaction?

Further Analysis of Inverted U-shaped Trend



1 **Bootstrap**

estimate confidence interval of peak CIIR

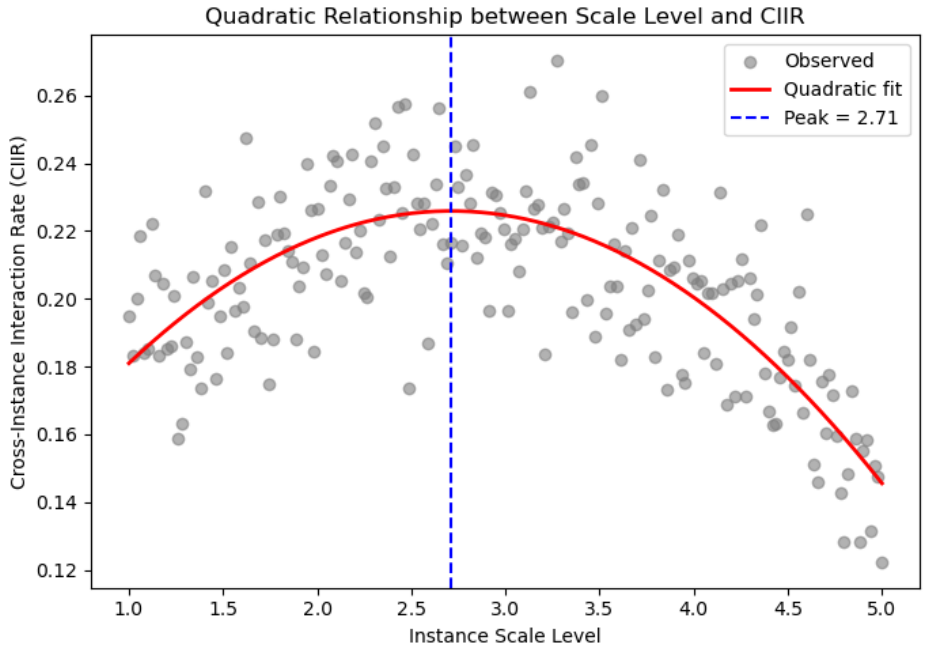
2 **Robustness check**

RQ1 Behavioral Motivation Layer: How does instance size affect cross-instance interaction?

Further Analysis of Inverted U-shaped Trend

Quadratic Regression Results

Coefficient	Value	t-value	p-value	Interpretation
β_0 (Intercept)	0.1131	12.365	<0.001	–
β_1 (Linear term)	0.0832	12.475	<0.001	–
β_2 (Quadratic term)	-0.0153	-14.008	<0.001	Negative and highly significant



$R^2 = 0.566$

model explains 56.6% of variance in CIIR

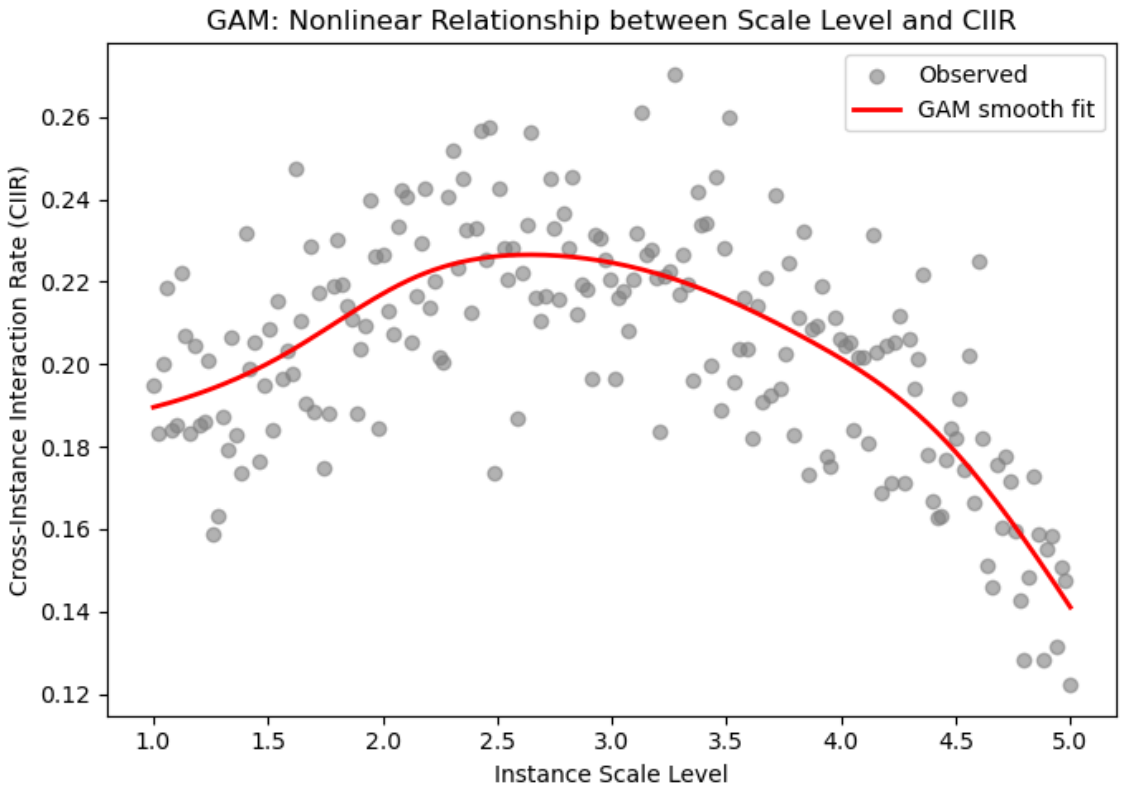
peak ≈ 2.71

CIIR reaches its maximum at level ≈ 2.71, and then decreases as the instance size continues to increase.

RQ1 Behavioral Motivation Layer: How does instance size affect cross-instance interaction?

Further Analysis of Inverted U-shaped Trend

GAM Results



The curve clearly shows an inverted U-shape, rising first and then falling.

Bootstrap Results

$$95\%Confidence\ Interval(\beta_2) = [-0.01756, -0{,}01333]$$

The entire confidence interval is negative and does not include 0, indicating that the direction of the quadratic term (negative) is consistently observed across resamples, and **the inverted U-shaped** relationship is **statistically robust.**

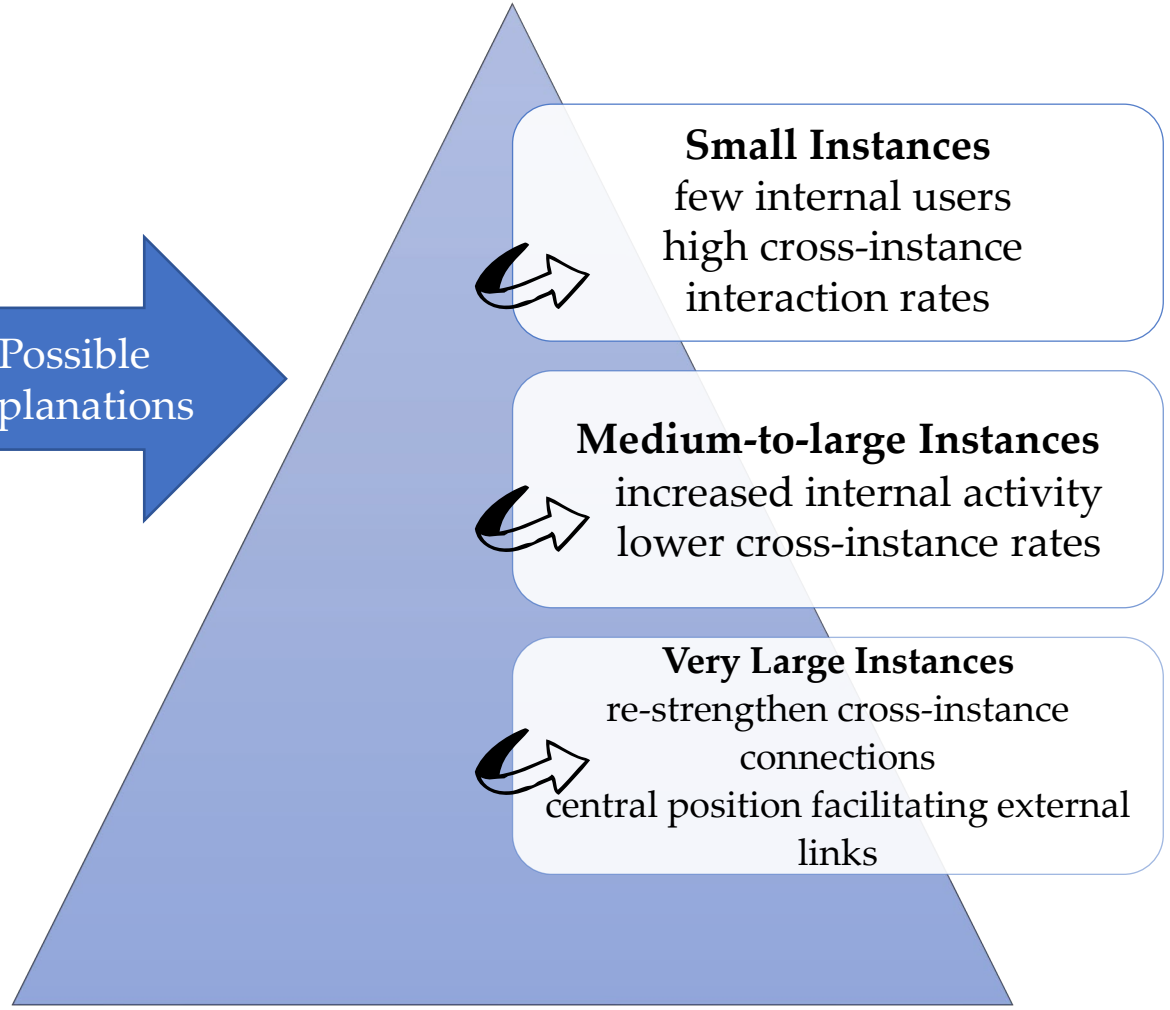
RQ1 Behavioral Motivation Layer: How does instance size affect cross-instance interaction?

Conclusions

Conclusions

- There is a significant and robust inverted U-shaped relationship between CIIR and instance size, indicating that CIIR is highest at medium instance size levels.
- As the instance size increases further, CIIR decreases.

Possible explanations



RQ2 Semantic Structure Layer: How Languages and Topics Jointly Shape the Semantic Community Structure

Analyze the impact of languages&topics on semantic community structure

- Count language&topic proportions within each community
- Identify the dominant language and topic
- Calculate language/topic entropy

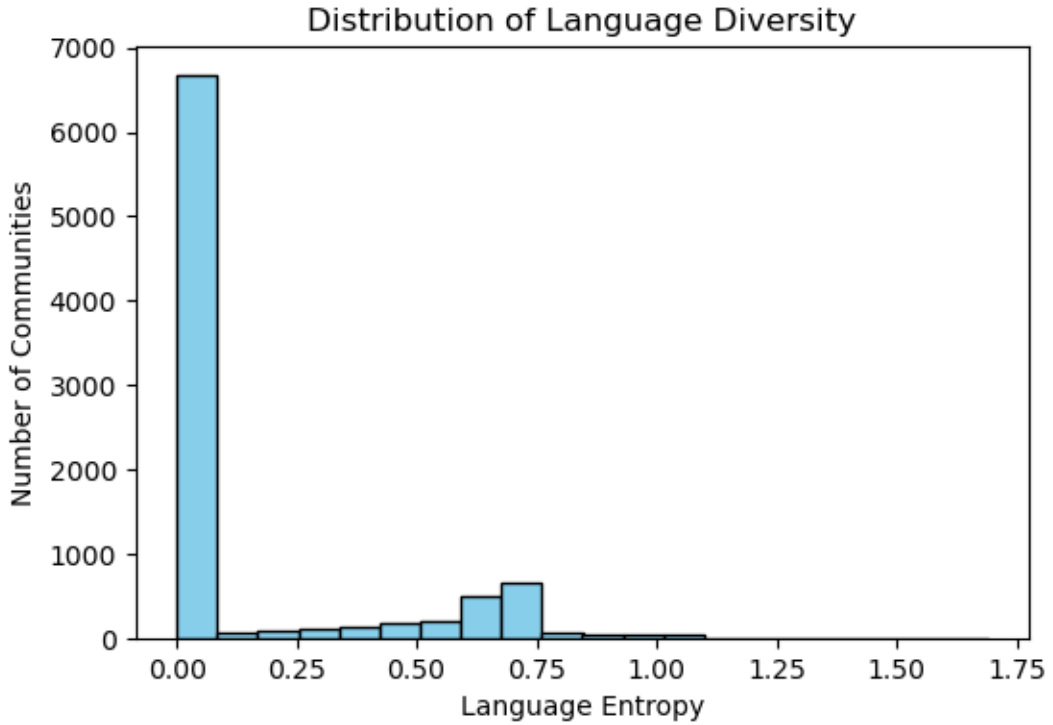
9,817 records

community_id	main_language	main_lang_prop	main_topic	main_topic_prop	lang_entropy	topic_entropy	community_size
0	zh-cn	0.521	主题_1	0.592	0.780	1.302	2179
1	en	0.531	游戏与娱乐	0.328	0.750	1.709	12767
2	en	0.641	游戏与娱乐	0.469	0.992	1.579	3836
3	en	0.585	主题_1	0.624	1.057	1.171	979
4	en	1.000	游戏与娱乐	1.000	0.000	0.000	2
5	en	0.566	游戏与娱乐	0.298	0.887	1.832	3447
6	en	1.000	游戏与娱乐	1.000	0.000	0.000	1
7	en	1.000	游戏与娱乐	1.000	0.000	0.000	1
8	en	0.922	游戏与娱乐	0.371	0.464	1.693	60164

RQ2 Semantic Structure Layer: How Languages and Topics Jointly Shape the Semantic Community Structure

Analyze the impact of languages&topics on semantic community structure

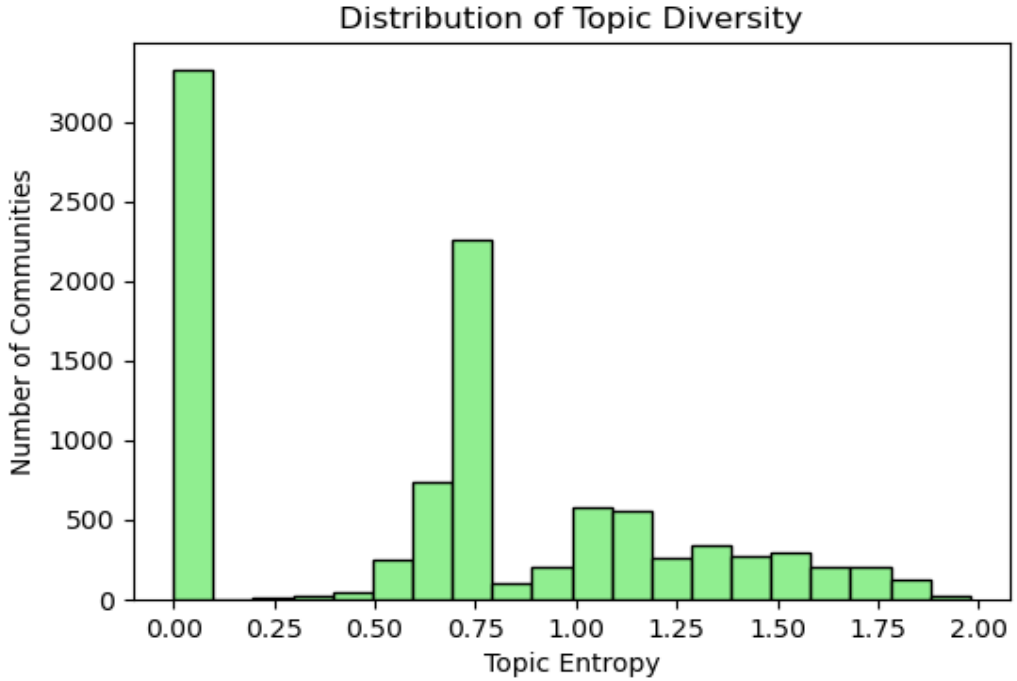
Language-driven Analysis



Average main language proportion: 0.92
Communities with main language proportion > 0.9:
6,992 / 9,817 (≈71%)
Average language entropy: 0.15

🗣Most communities are strongly dominated by a single language.

topic-driven Analysis



Average main topic proportion: 0.67
Communities with main topic proportion > 0.9:
3,348 / 9,817 (≈34%)
Average topic entropy: 0.64

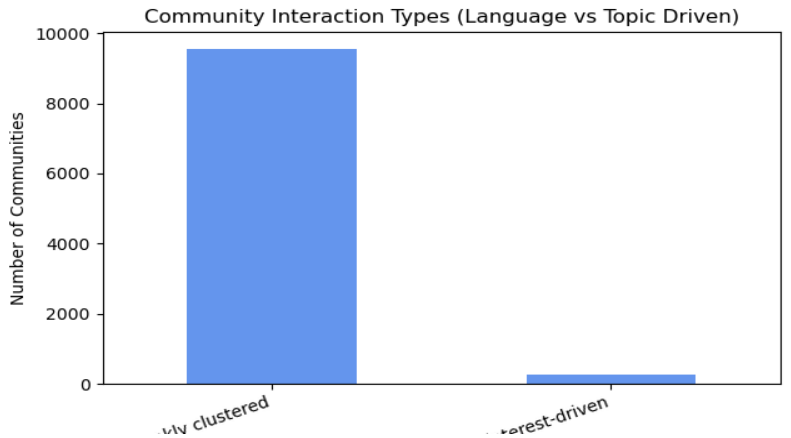
🧩 While most communities are language-homogeneous, they show noticeable diversity in topics.

RQ2 Semantic Structure Layer: How Languages and Topics Jointly Shape the Semantic Community Structure

Analyze the impact of languages&topics on semantic community structure

Type (divided by median language/topic entropy)	Count	Proportion	Interpretation
Both concentrated (Dual-driven)	0	**0**	-
Cross-lingual interest-driven	255	**2.6%**	These communities display clear topic concentration alongside high language diversity, indicating that when users interact around a shared interest, language differences become less important. In other words, common topics act as cross-language bridges and serve as a key driver of semantic community formation.
Language-concentrated / Topic-diverse (Topic-driven)	0	**0**	-
Both diverse (Weakly clustered)	9562	**97.4%**	Most communities exhibit high diversity in both language and topic, without forming distinct language- or topic-based clusters. This indicates that the overall semantic community structure is relatively loose, lacking stable linguistic or thematic centers.

Bar chart



RQ2 Semantic Structure Layer: How Languages and Topics Jointly Shape the Semantic Community Structure

Conclusions

- Language shapes the local cohesion of semantic communities, defining where linguistic boundaries tend to hold, whereas topics provide a cross-linguistic force that connects users beyond those boundaries.

- Together, the interplay between language and shared topics shapes how semantic communities emerge, fragment, or diffuse across the Mastodon network.

RQ3 : System Architecture Layer: Multidimensional Centrality and Optimal Scale Window

Distribution and Concentration Characteristics of Multidimensional Centrality

Network Construction

Remove intra-instance interactions

↓

Define active users: posts ≥ 1 & interactions ≥ 3

↓

Filter instances with ≥ 20 active users as nodes

↓

Construct network: 72 nodes, 1266 edges

Multidimensional Metric Definition

Structural Dimension

- Betweenness: Bridge role in network connectivity

Influence Dimension

- Katz: Influence considering attenuated long-range paths

Interaction Dimension （ε = 1e-6）

- Output Score: Per-capita weighted out-degree × diversity
- Input Score: Per-capita weighted in-degree × diversity

Output Score = (per capita weighted out-degree + ε)^0.6 × (per capita out-degree diversity + ε)^0.4

Input Score = (per capita weighted in-degree + ε)^0.6 × (per capita in-degree diversity + ε)^0.4

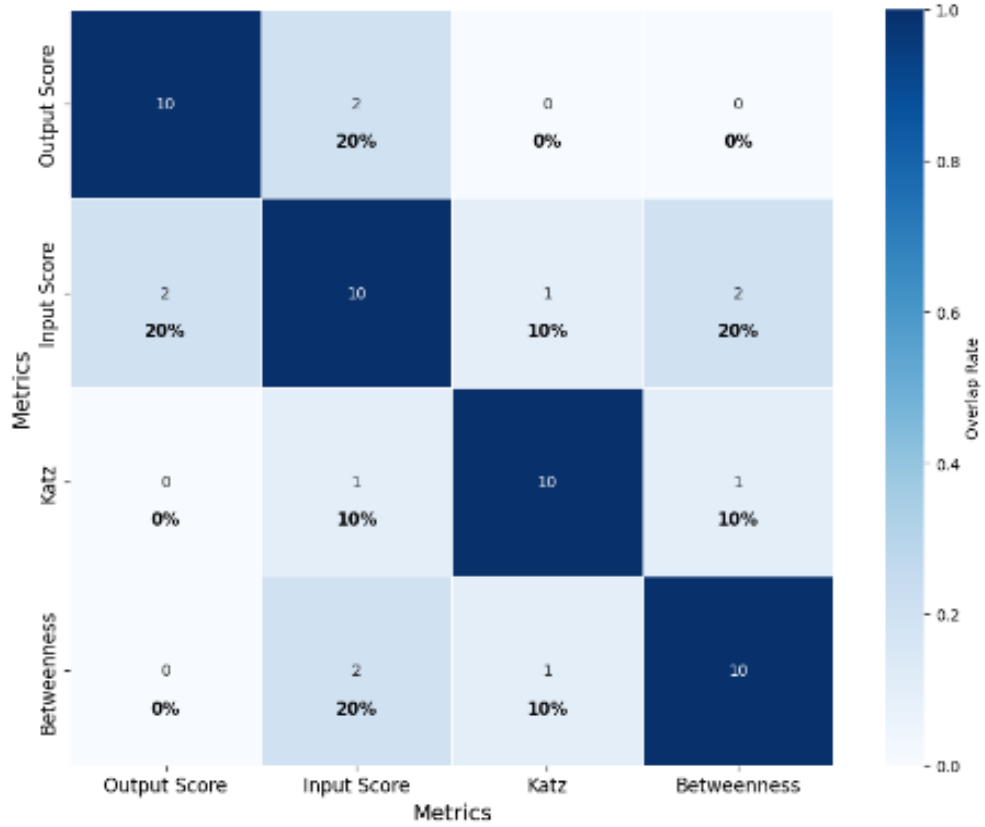
RQ3 : System Architecture Layer: Multidimensional Centrality and Optimal Scale Window

Distribution and Concentration Characteristics of Multidimensional Centrality

Core Differentiation Patterns

- Minimal overlap among Top-10 instances across metrics
- Different instances achieve prominence in distinct dimensions, demonstrating specialized roles and supporting multi-centered governance rather than single-point dominance

Figure 1 Top 10 Instances Overlap Analysis Across Different Metrics



RQ3 : System Architecture Layer: Multidimensional Centrality and Optimal Scale Window

Distribution and Concentration Characteristics of Multidimensional Centrality

🎯 Centralization Tendencies

- Heavy-tailed distributions patterns across all metrics
- the Gini coefficients of all four metrics exceed 0.5

🌐 Hub-and-Spoke Asymmetry

- Passive input scores significantly exceed active output scores
- Core instances dominate attention aggregation while non-core instances excel in cross-instance activity, revealing structural imbalances in federated interactions

Figure 3 Full Network Gini Coefficients Comparison (Selected Metrics)
(Darker blue indicates higher inequality)

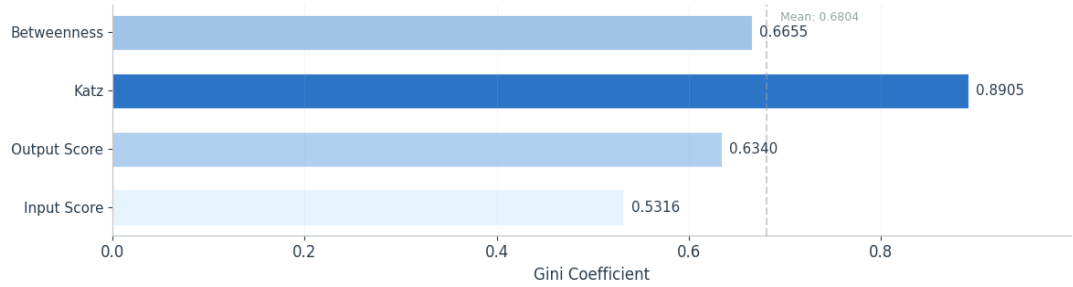
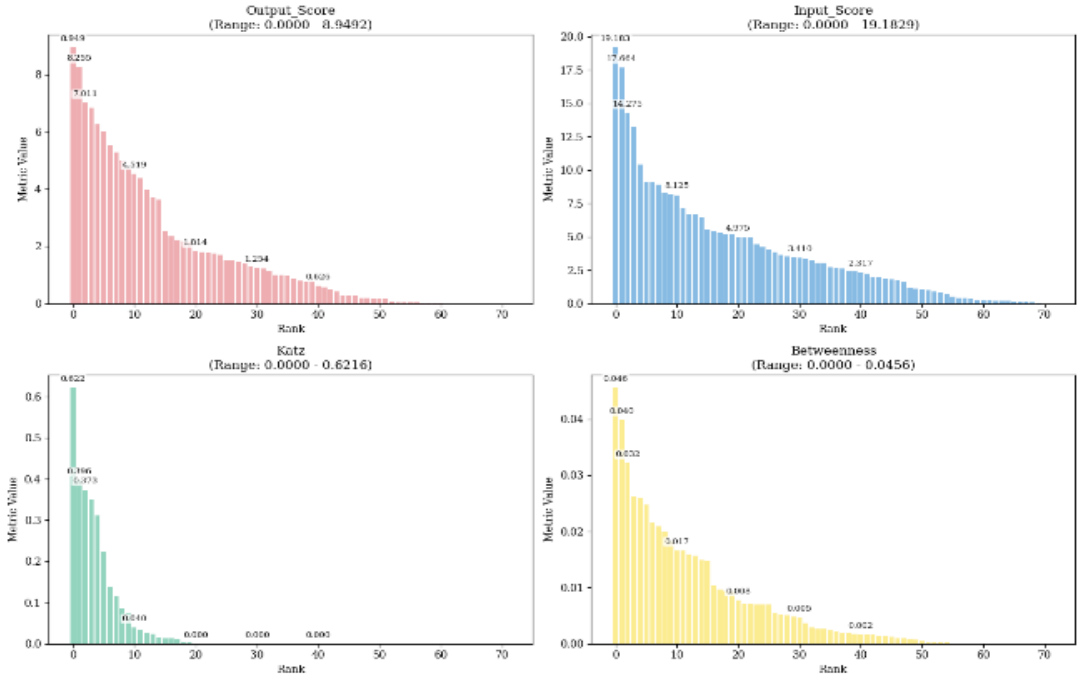


Figure 2 Original Scale Distribution of 4 Core Metrics



RQ3 : System Architecture Layer: Multidimensional Centrality and Optimal Scale Window

Core-Periphery Partition and Hierarchical Structure

Methodological Framework:

Across the four metrics, core networks are constructed individually—each comprising top-ranked instances with a dynamically scalable size ranging from 3 to 30—while the remaining instances form the non-core networks.

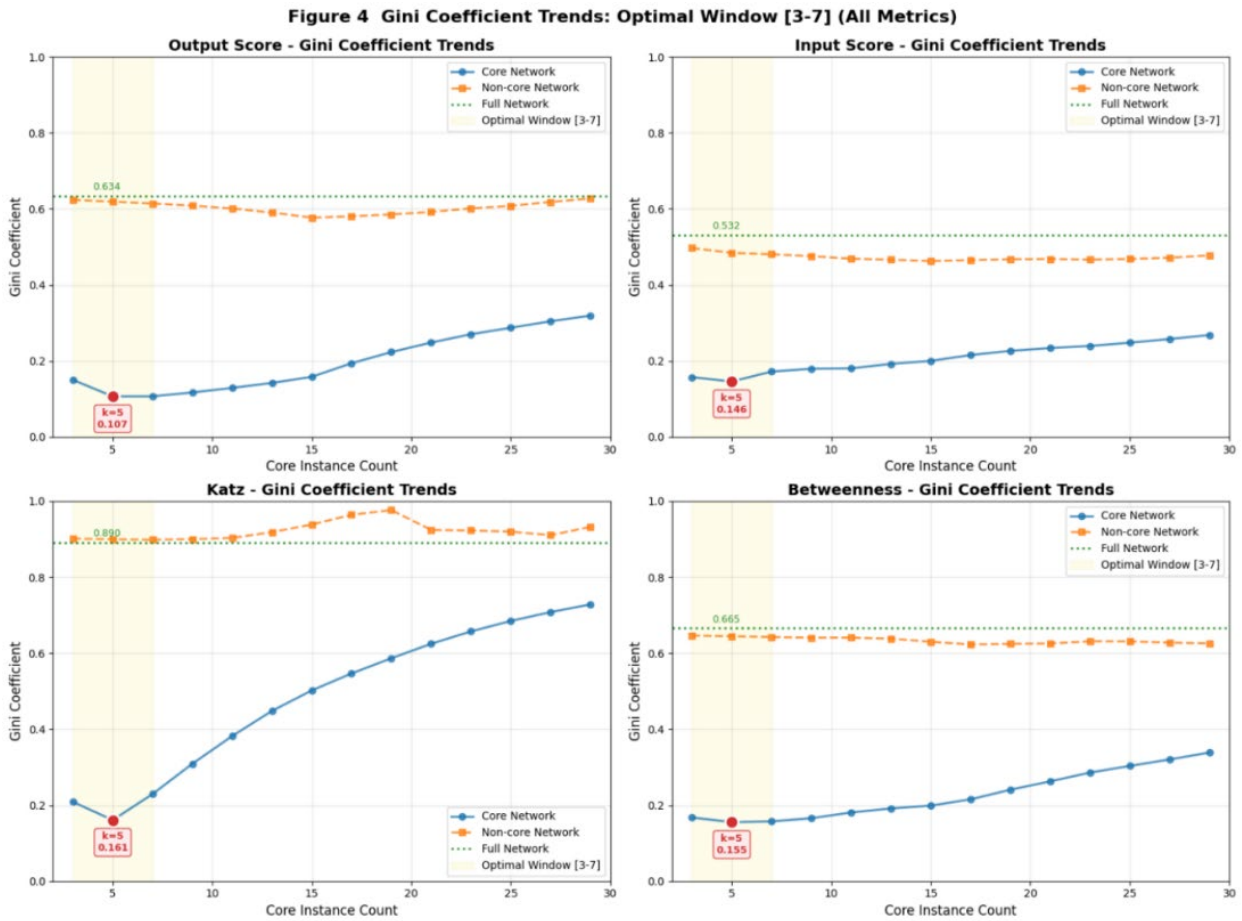
Subsequently, we calculate the Gini coefficient for each respective network using the corresponding metric.

RQ3 : System Architecture Layer: Multidimensional Centrality and Optimal Scale Window

Core-Periphery Partition and Hierarchical Structure

Optimal Scale Window:
All four metrics achieve peak decentralization within **Top 3-7 core instances**, with decay beyond this range
→ centralization reversion

Non-Core Stability Pattern:
All metrics show stable Gini coefficients in periphery networks
→ hard to obtain more balanced development opportunities due to changes in the scale of core instances



RQ3 : System Architecture Layer: Multidimensional Centrality and Optimal Scale Window

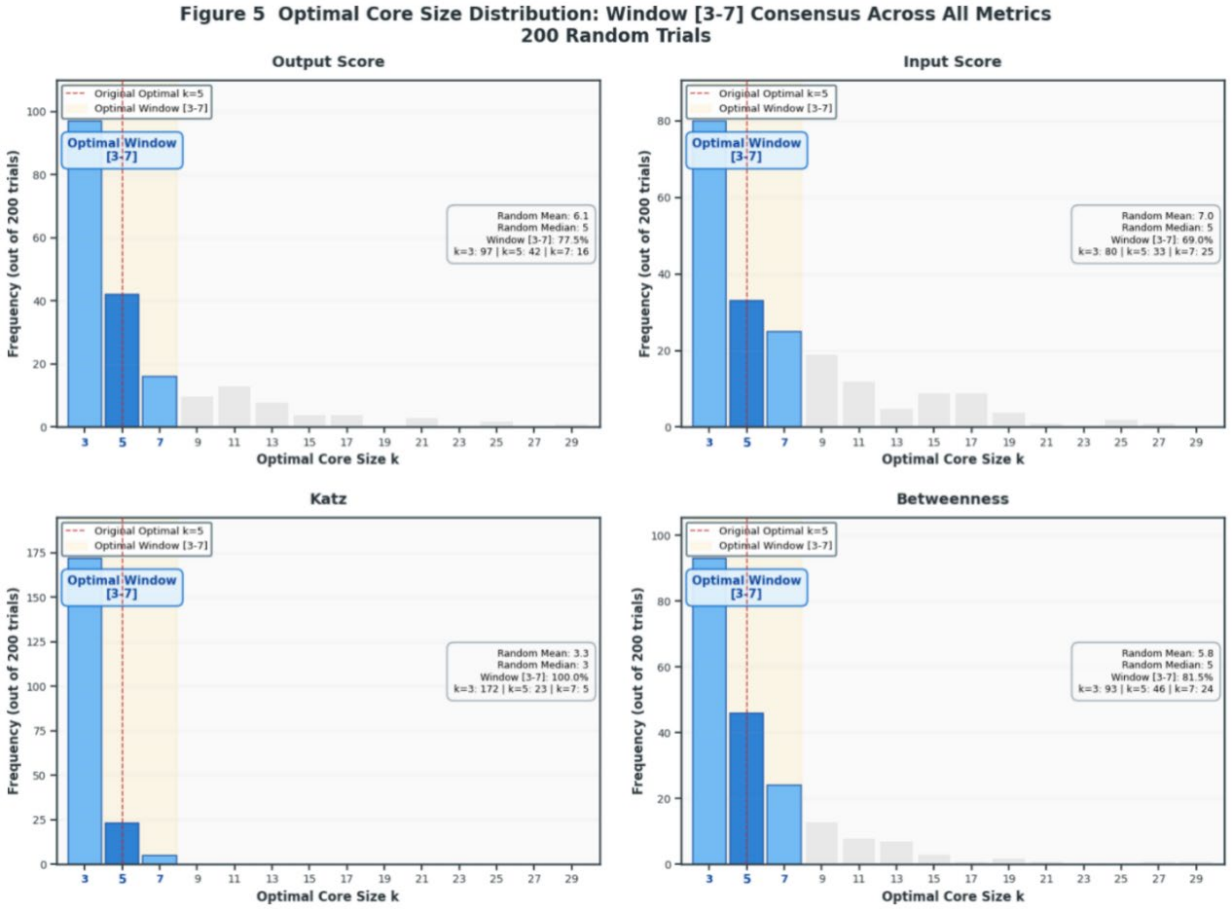
Robustness Verification of the Optimal Scale Window

Randomly shuffle the mapping relationships between the values of the metrics and the instances.

Identify the core size k*corresponding to the minimum Gini coefficient, and repeat this process 200 times.

The probability that k* falls within the interval [3,7] ranges from 69% to 100%.

The interval [3,7] is an endogenous steady-state interval.



RQ3 : System Architecture Layer: Multidimensional Centrality and Optimal Scale Window

Governance Implications for Decentralized Platforms

- Dynamically track Gini coefficient changes in both core and peripheral subnets. When the core size exceeds 7 and the core Gini coefficient rises without corresponding improvement in the peripheral Gini coefficient, the network can be deemed to have entered a "centralization reversion" state.

Establish a Decentralized Structural Health Monitoring and Early Warning Mechanism

- Provide necessary performance support for small and medium-sized instances through federated load balancing and service mirroring technologies. Default biases toward a small number of large-scale instances should be reduced in registration and content recommendation mechanisms.

Guide Mid-Tier Instances to Assume Active Network Roles via Traffic and Resource Diversion

- Implement diversity constraints and soft caps to incentivize reciprocal connections between core and peripheral instances, thereby reducing the positive feedback amplification effect caused by overly dense internal connections within the core;

Alleviate "Core Internal Circulation" and Enhance Structural Accessibility Between Core and Periphery

Conclusion

Limitation

- 13-day data window limits capture of federated networks' long-term evolution.
- Semantic labels from instance descriptions need improved refinement/consistency.
- Proposed governance strategies lack causal verification via intervention experiments.
- Decentralized governance framework requires deeper alignment with practical platform mechanisms.

Contribution

- Proposes a governance-oriented framework integrating cross-instance behavior, semantic communities, and multidimensional centrality.
- Provides empirical evidence for decentralization via ~1.36M Mastodon cross-instance interactions (13 days), identifying the [3–7] steady-state core size.
- Develops actionable metrics (active/passive scores, Gini coefficient, etc.) for federated platform decentralization monitoring and governance.

THANK YOU