Network, Covariates, or Both: A study on dynamic network data with covariates

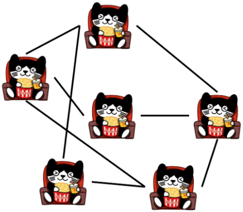
Bao Jiaqi

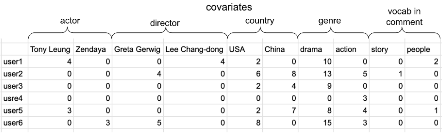
Supervisor:

Prof. Wang Wanjie

April 3, 2024

Static network data with covariates





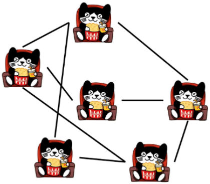
	covariates									
	actor		director		country		genre		vocab in comment	
	Tony Leung	Zendaya	Greta Gerwig	Lee Chang-dong	USA	China	drama	action	story	people
user1	4	0	0	4	2	0	10	0	0	2
user2	0	0	4	0	6	8	13	5	1	0
user3	0	0	0	0	2	4	9	0	0	0
usre4	0	0	0	0	0	0	0	3	0	0
user5	3	0	0	0	2	7	8	4	0	1
user6	0	3	5	0	8	0	15	3	0	0

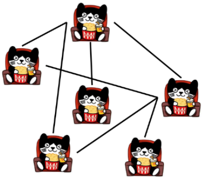
- Data is collected from Douban, the largest movie rating and review website in China.
- On a fixed date, the data comprises two main components: a static network and a static covariate matrix.
    - In the network, each node represents a Douban user. Two users are connected if they rated at least two of the same movies.
    - In the covariate matrix, covariates capture the personal details of the users, such as the actors in the movies they've rated and the specific vocabularies they've used in their movie comments.

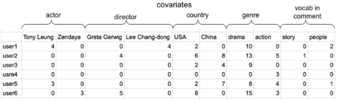
https://movie.douban.com/ Image from https://www.douban.com/group/topic/298502743/?_i=0751748cL-8KMg

Dynamic network data with covariates

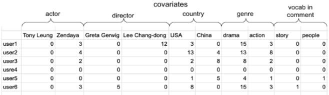
We collected data bi-weekly across 12 dates.







covariates										
	actor		director		country		genre		vocab in comment	
	Tony Leung	Zendaya	Greta Gerwig	Lee Chang-dong	USA	China	drama	action	story	people
user1	4	0	0	4	2	0	10	0	0	2
user2	0	0	4	0	6	8	13	5	1	0
user3	0	0	0	0	2	4	9	0	0	0
usre4	0	0	0	0	0	0	0	3	0	0
user5	3	0	0	0	2	7	8	4	0	1
user6	0	3	5	0	8	0	15	3	0	0



covariates										
	actor		director		country		genre		vocab in comment	
	Tony Leung	Zendaya	Greta Gerwig	Lee Chang-dong	USA	China	drama	action	story	people
user1	0	3	0	12	3	0	15	3	0	0
user2	0	4	0	0	13	4	13	8	0	0
user3	0	2	0	0	2	8	8	2	0	0
usre4	0	0	0	0	0	0	0	0	0	0
user5	0	0	0	0	1	5	4	1	0	1
user6	0	3	5	0	8	0	15	3	1	0

2023-04-14 ... 2023-09-15

Data description

- User and covariate selection:
  - A total of 2,206 active users, 3,832 popular covariates such as famous actors, directors, common movie genres, production countries and frequent vocabs in comments.
- For each date, with the user data from the preceding two weeks, we generate data in the following way:
  - Two users are connected in the network if they have rated at least two of the same movies.
  - The value of vocab covariate equals the frequency of a specific vocab in a user's comments.
  - The value of director, actor, genre and country covariate for a user equals the sum of rating scores (scale from 1 to 5) of all movies associated with a specific covariate rated by a user.

Objectives

1. Identify clusters of users with homogeneous **static covariates** and interpret their movie interest.
2. Explore the evolvement of user interest revealed by their dynamic cluster memberships in **dynamic covariates**.
3. Identify densely connected user clusters in **static network** and interpret their movie interest.
4. Identify user clusters in **static network with covariates** and interpret their movie interest.

Methods

- Data:
  - Actor and director covariate matrix on 2023-04-28, consisting of 1,880 users and 1,700 covariates.
  - Vocab covariate matrix on 2023-04-28, comprising 1,675 users and 1,765 covariates.
- Assumption [Hofmann, 1999]: In one covariate matrix, there are K distinct user clusters.
  - Each user i has a cluster membership vector π_i of length K, where $\pi_i(k)$ denotes the probability of user i belonging to cluster k. π_i represents user i's interest across different clusters.
  - Each cluster k has a distribution over the normalized covariates with a common mean.
- Goal: Estimate each user's cluster membership, as well as each cluster's distribution over covariates from both two covariate matrices.

Methods

Four algorithms:

1. Non-Negative Matrix Factorization (NMF) [Lee and Seung, 1999]
   - Decomposes the covariate matrix into components that represent underlying clusters.
   - Often results in extremely unbalanced clustering especially when K is limited.
2. Knowledge-guided NMF (KGNMF) [Chen et al., 2019]
   - Incorporates a penalty for large deviations in the estimated cluster's distribution over covariates, as compared to some pre-trained covariate embeddings.
   - Can only be applied to the vocab covariate matrix.
3. Latent Dirichlet Allocation (LDA) [Blei et al., 2003, Heinrich, 2005]
   - A Bayesian model.
   - Frequently encounters challenges when users mostly rate or use a small number of covariates.
4. Topic-SCORE (TSCORE) [Ke and Wang, 2022b]
   - A faster spectral approach with theoretical guarantees for the recovery of user clusters and their respective distributions over covariates.
   - Assumes all users rate or use a similar number of covariates.

Evaluation of actor and director covariates clustering

- NMF with 10 clusters attains the highest Quality.

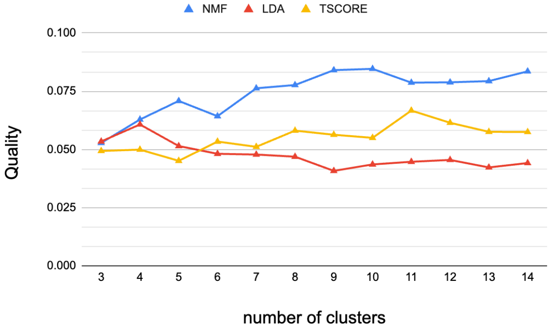


Figure: Higher Quality [Mimno et al., 2011, Lau et al., 2014, Dieng et al., 2020] suggests that the estimated distribution over covariates are more coherent within clusters and diverse between different clusters.

Interpretations of actor and director clusters

- Each user i is assigned to their most probable cluster: $\arg\max_{k} \hat{\pi}_i(k)$.
- We summarize the movie interest of each interpretable cluster based on its Top20 most probable actors and directors.
- We observe 5 major clusters, corresponding to 5 different regions.
- 3 large clusters are densely connected in the network on 04-28.

cluster interpretation	size	network density
Art films (France)	387	0.1014
Chinese movies	323	0.3690
American action and sci-fi movies	426	0.1409
Hong Kong movies	436	0.0360
European and American movies	202	0.0142
Harry Potter related	52	0.0301
Japanese anime	35	0.0941
Chinese classic TV series	19	0.1637

Evaluation of vocab covariates clustering

- KGNMF with $K = 7$ clusters exhibits the largest increase in Quality compared to $K - 2$.

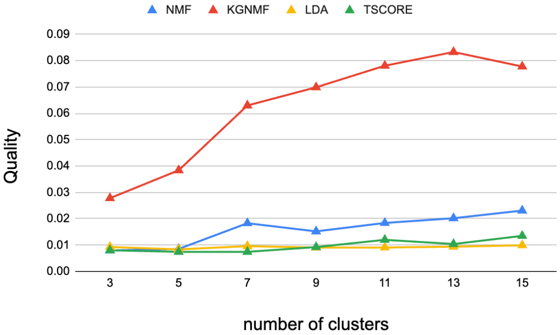


Figure: Higher Quality suggests that the estimated distribution over covariates are more coherent within clusters and diverse between different clusters.

Interpretations of vocab clusters

- We summarize the comment writing style of each interpretable cluster based on its Top20 most probable vocabs.
- There are 3 major movie comment writing styles.
- All vocab clusters are loosely connected in the network on 04-28.

cluster interpretation	size	network density	Top6 most probable vocab covariates
common words	1160	0.0579	电影 (movie), 人 (people), 故事 (story), 女主 (female lead), 生活 (life), 导演 (director)
film critiques	310	0.0529	影片 (film), 结构 (structure), 影像 (image), 敘事 (narrative), 情感 (feeling), 故事片 (drama movies)
technical aspects	183	0.0315	同名 (homonyms), 战争 (war), 小说 (novel), 创作 (creation), 代表作 (representative works), 整体 (whole)
writing	10	0.0444	本作 (original work), 单元 (unit), 人物 (character), 三部曲 (trilogy), 动作 (movement), 空间 (space)
symbols	8	0	镜子 (mirror), 儿子 (son), 镜像 (mirrored), 怪兽 (monster), 黑人 (black people), 战争 (war)
art	4	0	资料馆 (film archive), 权力 (power), 影像 (image), 女人 (women), 矛盾 (conflict), 本质 (essence)

Comparison of actor and director clusters with vocab clusters

- The two clusterings are weakly correlated, but there is a stronger alignment between "art films" and "film critiques".

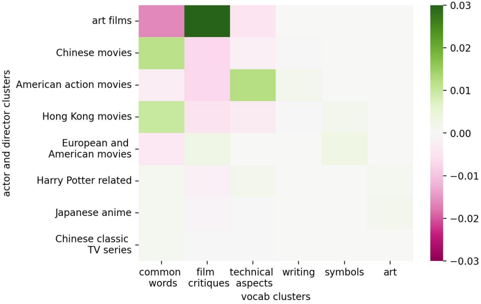


Figure: A higher value [Danon et al., 2005] or darker green color at the (k_1, k_2)-th entry suggests that actor and director cluster k_1 and vocab cluster k_2 are more aligned in terms of their mutual members.

Methods

- Data: All 12 actor and director covariate matrices collected from 2023-04-14 to 2023-09-15. It encompasses a total of 798 users and 1,150 covariates.
- Assumption: There are K dynamic clusters. On each date t, each user has a cluster membership vector $\pi_i^{(t)}$ and each cluster has a distribution over covariates.
- Goal: On each date, we aim to estimate each user's cluster membership and each cluster's distribution over covariates.

Methods

Topic Tracking Model (TTM) [Iwata et al., 2009]

- A dynamic generalization of LDA
- For a given date, the estimation of a user's cluster membership, along with a cluster's distribution over covariates, relies on
    - the covariate data observed on this date
    - the estimates obtained from the preceding dates.

Dynamic popularity of clusters

- We choose $K = 5$ dynamic clusters.
- On each date t, each user i is assigned to their most probable cluster $\arg\max_k \hat{\pi}_i^{(t)}(k)$

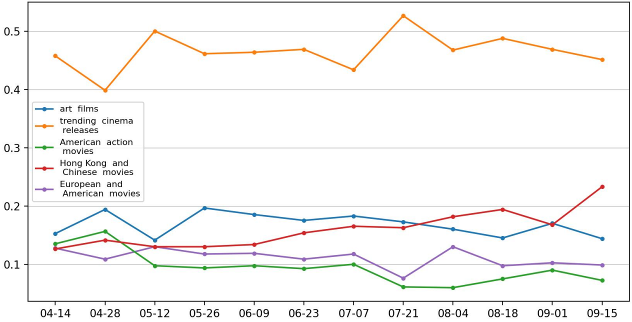


Figure: The proportion of users in each dynamic cluster is plotted on each date. Each line illustrates the changing popularity of the corresponding cluster over time.

User classification based on their time-varying interest

- Each user i has a *trace* s_i of length 12, where $s_i(t)$ represents the cluster label of user i on date t. s_i represents the evolving interest of user i.
- We construct a user-user weighted network, where the edge weight between user i and j depends on the proportion of the same entries in s_i and s_j.
- We perform spectral clustering [Donath and Hoffman, 1973, Rohe et al., 2011] on the network.
- For each user group $\mathcal{G}_\ell$ discovered, extract a small set of representative user traces s_{i^*}'s, such that for many $i \in \mathcal{G}_\ell$, the edge weight between user i and i^* is high.

group	size	network density	representative traces
$\mathcal{G}_1$	271	0.5622	2 2 2 2 2 2 2 2 2 2 2 2
$\mathcal{G}_2$	178	0.1777	2 2 2 4 4 4 4 4 2 4 4 4 2 2 2 2 2 2 4 2 4 2 2 4 4 4 2 4 3 5 4 4 4 4 4 4 2 2 4 2 4 4 4 2 2 4 2 4
$\mathcal{G}_3$	123	0.2677	1 1 2 1 1 1 1 1 1 1 2 2
$\mathcal{G}_4$	226	0.1215	1 1 5 2 1 2 2 2 2 2 2 2 ...

Table: 1: art films 2: trending cinema releases 3: American action movies 4: Hong Kong and Chinese movies 5: European and American movies

User classification based on their time-varying interest

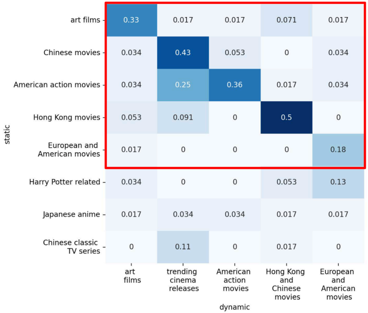
- Each user i has a *trace* s_i of length 12, where $s_i(t)$ represents the cluster label of user i on date t. s_i represents the evolving interest of user i.
- We construct a user-user weighted network, where the edge weight between user i and j depends on the proportion of the same entries in s_i and s_j.
- We perform spectral clustering [Donath and Hoffman, 1973, Rohe et al., 2011] on the network.
- For each user group $\mathcal{G}_\ell$ discovered, extract a small set of representative user traces s_{i^*}'s, such that for many $i \in \mathcal{G}_\ell$, the edge weight between user i and i^* is high.

group	size	network density	representative traces
$\mathcal{G}_1$	271	0.5622	2 2 2 2 2 2 2 2 2 2 2 2
$\mathcal{G}_2$	178	0.1777	2 2 2 4 4 4 4 4 2 4 4 4 2 2 2 2 2 2 4 2 4 2 2 4 4 4 2 4 3 5 4 4 4 4 4 4 2 2 4 2 4 4 4 2 2 4 2 4
$\mathcal{G}_3$	123	0.2677	1 1 2 1 1 1 1 1 1 1 2 2
$\mathcal{G}_4$	226	0.1215	1 1 5 2 1 2 2 2 2 2 2 2 ...

Table: 1: art films 2: trending cinema releases 3: American action movies 4: Hong Kong and Chinese movies 5: European and American movies

Comparison of static and dynamic user clusters on 04-28

- 5 static and dynamic cluster pairs display similar estimated distribution over covariates, indicating an alignment in their movie interests.
- The long-term interest behind the static Chinese movie cluster is more potentially towards recently released trending movies.



static \ dynamic	art films	trending cinema releases	American action movies	Hong Kong and Chinese movies	European and American movies
art films	0.33	0.017	0.017	0.071	0.017
Chinese movies	0.034	0.43	0.053	0	0.034
American action movies	0.034	0.25	0.36	0.017	0.034
Hong Kong movies	0.053	0.091	0	0.5	0
European and American movies	0.017	0	0	0	0.18
Harry Potter related	0.034	0	0	0.053	0.13
Japanese anime	0.017	0.034	0.034	0.017	0.017
Chinese classic TV series	0	0.11	0	0.017	0

Figure: The (k, ℓ)-th entry of the matrix represents the Jaccard similarity between the Top20 most probable covariates of static cluster k and dynamic cluster ℓ on 04-28.

Comparison of static and dynamic user clusters on 04-28

- Interests in “trending cinema releases” and “art films” demonstrate more persistence and are easy to predict.
- Interests in “American action movies”, “Hong Kong movies” and “European and American movies” tend to be more occasional and unpredictable.

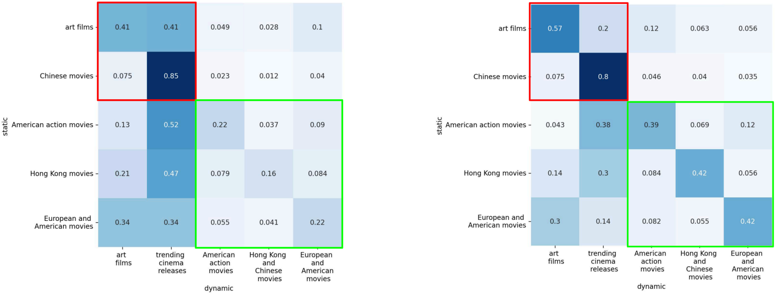


Figure: Left: The (k,ℓ)-th entry represents $|\mathcal{S}_k \cap \mathcal{D}_\ell^{(t-1)}|/|\mathcal{S}_k|$, where $\mathcal{D}_\ell^{(t-1)}$ denotes each dynamic cluster 04-14 and $\mathcal{S}_k$ denotes each static cluster. Right: The (k,ℓ)-th entry represents $|\mathcal{S}_k \cap \mathcal{D}_\ell^{(t)}|/|\mathcal{S}_k|$, where $\mathcal{D}_\ell^{(t)}$ denotes each dynamic cluster on 04-28.

Methods

- Data: We will cluster the largest connected component of the network collected on 2023-04-28. This component, which consists of 1,672 users, serves as the core of the network.
- Assumption [Karrer and Newman, 2011]: There are K user clusters, and the likelihood of connection between users i and j depends on their respective cluster labels and degree heterogeneity, θ_i and θ_j.
  - Connections within clusters are more likely compared to connections across different clusters.
  - Each user is associated with a degree heterogeneity parameter θ_i, where a higher value indicates a greater inherent likelihood for user i to connect with others.
- Goal: Recover each user's cluster label from the network.

Methods

Two spectral algorithms:

1. Spectral Clustering On Ratios-of-Eigenvectors (SCORE) [Jin, 2015, Ke and Jin, 2023] is able to recover cluster structures in a network with high average user degree and moderate degree heterogeneity.
2. Regularized spectral clustering (RSC) [Qin and Rohe, 2013, Joseph and Yu, 2016, Ke and Wang, 2022a] can recover clusters of users with high degrees under relatively severe degree heterogeneity, but struggles with separating multiple clusters of users with low degrees.

- The degree of a user equals the total number of connections they have with other users.
- The level of degree heterogeneity can be assessed by examining the distribution of θ_i's.

Evaluation of network clustering

- RSC identifies clusters with stronger connection and higher coherence or homogeneity in their covariate cluster labels.

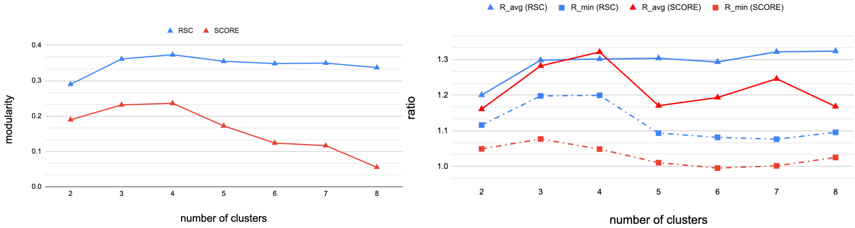


Figure: Left: Higher modularity [Newman, 2006, Zhao et al., 2012] indicates network clusters with stronger connections. Right: Higher ratios R_{avg} and $R_{\min}$ [Bordino et al., 2010, Yan et al., 2013] indicate greater coherence in the estimated actor and director covariate cluster labels within network clusters, and increased variation across different network clusters.

Interpretation of network clusters

- We choose RSC with 8 clusters.
- We are able to identify 3 covariate clusters: "art films" as $\mathcal{N}_1$, "Chinese movies" as $\mathcal{N}_2$, and "American action movies" as $\mathcal{N}_3$ using network only.
- Reveals 3 new clusters: $\mathcal{N}_4$, $\mathcal{N}_5$ and $\mathcal{N}_6$ display interests in two distinct covariate clusters.

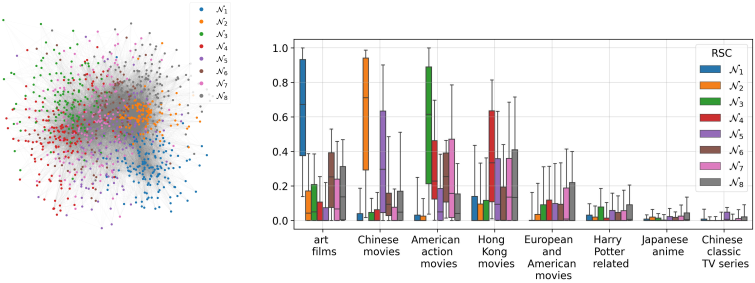


Figure: Left: Estimated network cluster labels of users with degree ≥ 10. Right: The box associated with the estimated network cluster $\mathcal{N}_\ell$ and actor and director cluster k represents the set $\{\hat{\pi}_i(k) | i \in \mathcal{N}_\ell\}$, where $\hat{\pi}_i(k)$ denotes the estimated probability for user i within covariate cluster k.

Methods

- Data: The network with actor and director covariates collected on 04-28. It comprises a total of 2,052 users and 1,700 covariates.
- Goal: We aim to recover each user's network cluster label by incorporating covariates, instead of recovering user's covariate cluster labels.

Methods

Three spectral algorithms:

1. Covariate assisted spectral clustering (CASC) [Binkiewicz et al., 2017] incorporates covariates into RSC. This approach, however, assumes the perfect alignment between user clusters in covariates and networks, and struggles to cluster low degree users similar to the limitations of RSC.
2. GraphText [Zhang et al., 2018] refines CASC, utilizing a sparsity penalty to select covariates that align with the network structure.
3. Network adjusted covariates (NAC) [Hu and Wang, 2024] can recover the network cluster labels of users with low degrees if their covariate cluster labels align with their network labels.

Evaluation of network with covariates clustering

- GraphText consistently obtains the lowest R_{min}, indicating the discovery of some cluster with incoherent interest in covariates.

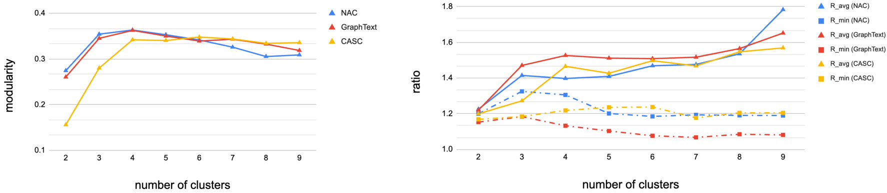


Figure: Left: Higher modularity indicates stronger network cluster structures. Right: Higher ratios R_{avg} and R_{min} indicate greater coherence in the estimated actor and director cluster labels within the identified clusters, and increased variation across different clusters.

Interpretation of clusters

- We select the number of clusters to be 8 for both CASC and NAC.

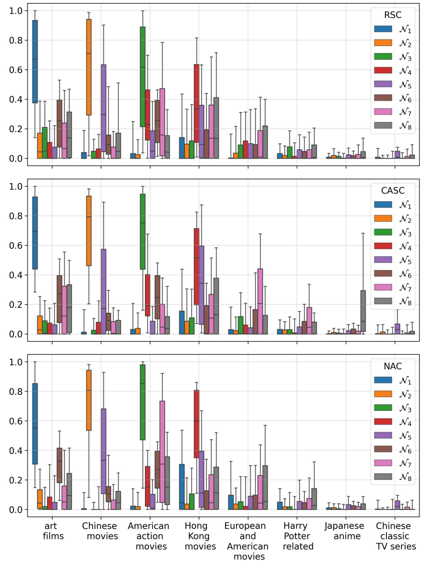


Figure: The box associated with the estimated cluster $\mathcal{N}_\ell$ and covariate cluster k represents the set $\{\hat{\pi}_i(k) | i \in \mathcal{N}_\ell\}$, where $\hat{\pi}_i(k)$ denotes the estimated probability for user i within covariate cluster k.

- $\mathcal{N}_1$, $\mathcal{N}_2$, $\mathcal{N}_3$, $\mathcal{N}_5$, and $\mathcal{N}_6$, estimated by all 3 methods using either network or network with covariates, display shared interests.
- $\mathcal{N}_4$ estimated from the network with covariates, exhibits a stronger interest in Hong Kong movies compared to its corresponding cluster estimated from the network alone.

Cluster attention toward production country and genre

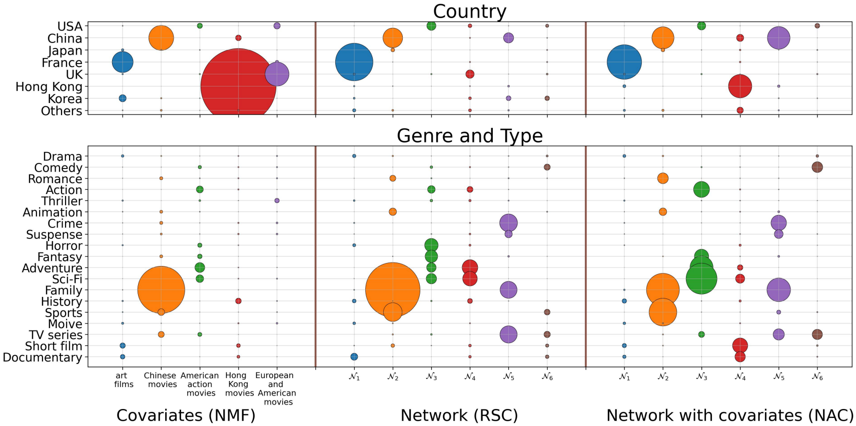


Figure: The size of the circle at the (j,k)-th entry of the matrix depends on Ψ_{kj}, which measures the attention [Zhang et al., 2018] that cluster ω_k pays to covariate j. We calculate $\Psi_{kj} = (\sum_{i\in\omega_k} \widetilde{X}_{ij}/|\omega_k|)/(\sum_i \widetilde{X}_{ij}/n)$, where X denotes the country, genre and type covariate matrix on 04-28, n denotes the number of users and $\widetilde{X}_{ij} = X_{ij}/(\sum_j X_{ij})$.

Conclusion

We construct a dynamic user network with covariates. We apply a range of state-of-the-art techniques, our findings are summarized below.

1. Static covariates:
    - 5 major region-centered covariate clusters: “art films”, “Chinese movies”, “American action movies”, “Hong Kong movies”, and “European and American movies”.
    - The “art film” cluster writes more professional movie critiques in their comments.
2. Dynamic covariates:
    - The long term interest of the static “Chinese movie” cluster tends to be “trending cinema releases”.
    - The “trending cinema releases” cluster consistently remains the most popular. The “American action movie” cluster is shrinking, while the “Hong Kong movie” cluster is growing.
    - Interests in “art films” and “trending cinema releases” exhibit higher persistence and stability.

Conclusion

3. Static network:
    - 3 densely connected "art films", "Chinese movies", "American action movies" covariate clusters are again discovered.
    - 3 new network clusters reflecting users' movie genre interest. One shows interest in both "Hong Kong and American action movies", another focuses on "Chinese crime TV series", and a third cluster prefers both "Hollywood commercial and art movies".
4. Static network with covariates:
    - 3 covariate or network clusters: "Art films", "Chinese movies", "American action movies", and 2 network clusters: "Chinese crime TV series", and "Hollywood commercial and art movies" are again identified.
    - Suggests that the "Hong Kong and American action movies" network cluster has a stronger inclination towards Hong Kong movies.

Thank You!

Binkiewicz, N., Vogelstein, J. T., and Rohe, K. (2017).
Covariate-assisted spectral clustering.
Biometrika, 104(2):361–377.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003).
Latent dirichlet allocation.
Journal of machine Learning research, 3(Jan):993–1022.

Bordino, I., Castillo, C., Donato, D., and Gionis, A. (2010).
Query similarity by projecting the query-flow graph.
In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 515–522.

Chen, Y., Zhang, H., Liu, R., Ye, Z., and Lin, J. (2019).
Experimental explorations on short text topic mining between lda and nmf based schemes.
Knowledge-Based Systems, 163:1–13.

Danon, L., Diaz-Guilera, A., Duch, J., and Arenas, A. (2005).
Comparing community structure identification.
Journal of statistical mechanics: Theory and experiment, 2005(09):P09008.

Dieng, A. B., Ruiz, F. J., and Blei, D. M. (2020).
Topic modeling in embedding spaces.
Transactions of the Association for Computational Linguistics, 8:439–453.

Donath, W. E. and Hoffman, A. J. (1973).
Lower bounds for the partitioning of graphs.
IBM Journal of Research and Development, 17(5):420–425.

Heinrich, G. (2005).
Parameter estimation for text analysis.
Technical report, Citeseer.

Hofmann, T. (1999).
Probabilistic latent semantic indexing.
In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57.

Hu, Y. and Wang, W. (2024).
Network-adjusted covariates for community detection.
Biometrika, page asae011.

Iwata, T., Watanabe, S., Yamada, T., and Ueda, N. (2009).
Topic tracking model for analyzing consumer purchase behavior.
In *Twenty-First international joint conference on artificial intelligence*.

Jin, J. (2015).
Fast network community detection by score.
The Annals of Statistics, 43(1):57–89.

Joseph, A. and Yu, B. (2016).
Impact of regularization on spectral clustering.
The Annals of Statistics, 44:1765–1791.

Karrer, B. and Newman, M. E. (2011).
Stochastic blockmodels and community structure in networks.
Physical review E, 83(1):016107.

Ke, Z. T. and Jin, J. (2023).
Special invited paper: The score normalization, especially for heterogeneous network and text data.
Stat, 12(1):e545.

Ke, Z. T. and Wang, J. (2022a).
Optimal network membership estimation under severe degree heterogeneity.
arXiv preprint arXiv:2204.12087.

Ke, Z. T. and Wang, M. (2022b).
Using svd for topic modeling.
Journal of the American Statistical Association, pages 1–16.

Lau, J. H., Newman, D., and Baldwin, T. (2014).
Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality.
In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539.

Lee, D. D. and Seung, H. S. (1999).
Learning the parts of objects by non-negative matrix factorization.
Nature, 401(6755):788–791.

Mimno, D., Wallach, H., Talley, E., Leenders, M., and McCallum, A. (2011).
Optimizing semantic coherence in topic models.
In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 262–272.

Newman, M. E. (2006).
Modularity and community structure in networks.
Proceedings of the national academy of sciences, 103(23):8577–8582.

Qin, T. and Rohe, K. (2013).
Regularized spectral clustering under the degree-corrected stochastic blockmodel.
Advances in neural information processing systems, 26.

Rohe, K., Chatterjee, S., and Yu, B. (2011).
Spectral clustering and the high-dimensional stochastic blockmodel.
The Annals of Statistics, 39(4):1878–1915.

Yan, X., Guo, J., Lan, Y., and Cheng, X. (2013).
A biterm topic model for short texts.
In *Proceedings of the 22nd international conference on World Wide Web*, pages 1445–1456.

Zhang, Y., Poux-Berthe, M., Wells, C., Koc-Michalska, K., and Rohe, K. (2018).
Discovering political topics in facebook discussion threads with graph contextualization.
Annals of Applied Statistics, 12(2):1096–1123.

Zhao, Y., Levina, E., and Zhu, J. (2012).
Consistency of community detection in networks under degree-corrected stochastic block models.
The Annals of Statistics, 40(4):2266–2292.