AFRL-AFOSR-JP-TR-2023

Linking Online Engagement to Pandemic Responses

Xie, Lexing
AUSTRALIAN NATIONAL UNIVERSITY RESEARCH OFFICE ACTON (AUSTRALIA)
10C EAST RD
ACTON, ,
AU

01/27/2023
Final Technical Report

Air Force Research Laboratory
Air Force Office of Scientific Research
Asian Office of Aerospace Research and Development
Unit 45002, APO AP 96338-5002

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
20230127	Final	START DATE 20200930	END DATE 20220929

4. TITLE AND SUBTITLE
Linking Online Engagement to Pandemic Responses

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
	FA2386-20-1-4064	61102F
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)
Lexing Xie

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
AUSTRALIAN NATIONAL UNIVERSITY RESEARCH OFFICE ACTON (AUSTRALIA)
10C EAST RD
ACTON
AU

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AOARD UNIT 45002 APO AP 96338-5002	AFRL/AFOSR IOA	AFRL-AFOSR-JP-TR-2023-0057

12. DISTRIBUTION/AVAILABILITY STATEMENT
A Distribution Unlimited: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This project aimed to uncover the mathematical underpinnings of online attention market, design computational methods for making attention allocation stable and fair, enable applications where content producers, consumers, platforms and regulators could each play a role in influencing attention allocation towards socially beneficial goals. The new knowledge will enable organizations and individuals to foster safer and more trustworthy online spaces for the public, and benefit citizens through the minimization of harm and misinformation when they are online.
The project was highly successful and produced 3 very high quality papers which will be of significant value in the field. Of particular interest is their analysis of the socio-political effects of social media. The researchers clearly dedicated good people and and large amount of manpower to the effort. I would definitely recommend them for any future funding efforts.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	c. THIS PAGE		
U	U	U	SAR	66

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
GEOFFREY ANDERSEN	

AOARD/AFOSR FA2386-20-1-4064 – Final Report

Linking Online Engagement to Pandemic Responses**:**
Measuring Media Interactions in Pandemic Response

Principle investigator:
Lexing Xie, lexing.xie@anu.edu.au
Key personnel:
Alasdair Tran (ANU) alasdair.tran@anu.edu.au
Siqi Wu (University of Michigan and ANU) siqi.wu@anu.edu.au
Jooyoung Lee (Currently at University Techonology Sydney)
Research students:
Cai Yang, Haiqing Zhu
Organization:
Computational Media Lab: http://cm.cecs.anu.edu.au
School of Computing
The Australian National University, Acton, ACT 2601, Australia.

Collaborator (via AFSOR FA2386-20-1-4065)
Yu-Ru Lin yurulin@pitt.edu
School of Computing and Information,
University of Pittsburgh, Pittsburgh, PA 15260, USA.
Computational Social Dynamics Lab: https://picsolab.github.io
Period of Performance:
Sept 1 2020 – Aug 31 2022

Project highlights

This project breaks new grounds on three different fronts – in the *theory*, *algorithms* and large-scale *measurement studies* of online attention. The new advances are enabled but significant expands those made in our prior AFOSR projects: *The Anatomy of Social Media Popularity* (2015-2018, `#1514018`) and *Linking Online Attention to Measurable Actions* (2019-2021, `#1914078`).

On the **theory** of online attention, we propose a novel view of attention allocation via computational economics and mathematical optimization. Specifically, we examine the attention allocation of many users on a set of digital goods with infinite supply. Such dynamic is important in shaping processes and outcomes in society, from trending items in entertainment, collective knowledge creation, to election outcomes. The outcomes of online cultural markets are susceptible to intricate social influence dynamics, particularly so when the community comprises consumers with heterogeneous interests. This has made formal analysis of these markets improbable. In this work, we establish robust connections between influence dynamics and optimization processes for the first time. In trial-offer markets where the consumer preferences are modelled by multinomial logit, we show that the proportional-response-esque influence dynamic is equivalent to stochastic mirror descent on a convex objective function, thus leading to a stable and predictable outcome. When all consumers are homogeneous, the objective function has a natural interpretation as a weighted sum of efficiency and diversity of the culture market. In simulations driven by a large-scale recommender system, we observe that ranking strategies aligned with the underlying heterogeneous

preferences achieve higher efficiency and diversity.

On **algorithms** that explain and forecast online attention, we propose Radflow, a new, robust, and efficient model for large networks of time series. Radflow embodies three key ideas: a recurrent neural network to obtain node embeddings that depend on time, the aggregation of the flow of influence from neighboring nodes with multi-head attention, and the multi-layer decomposition of time series. On real-world datasets ranging from a few hundred to a few hundred thousand nodes, we observe that Radflow variants are the best performing model across a wide range of settings. The recurrent component in Radflow also outperforms N-BEATS, the state-of-the-art time series model. We show that Radflow can learn different trends and seasonal patterns, that it is robust to missing nodes and edges, and that correlated temporal patterns among network neighbors reflect influence strength.

On **measurements**, we present a large-scale cross-country study for characterizing the distribution of audience political leanings for different media outlets. The methodological components of these new measurements include high-fidelity Twitter streams; highly accurate user geolocations; mapping global users onto one coherent axis of ideological leaning; and comprehensive validation of the estimated new media bias scores. We apply the resulting metrics to many countries, covering many countries in the well-studied Global North, and those in the under-studied Global South. These data-driven tools help answer questions such as: Are US readers of the *cnn* on average more left-leaning than those of *theguardian*? Does *reuters* reach an ideologically balanced audience base in US, Argentina, and Philippines? As a preview, we find mixed results for these two questions. And our methods allow us to extend the questions to other media and countries. We expect that such answers would inform activist groups, and media watchdog organizations about country-specific media audience profiles, inform media outlets of their effectiveness and audience base across different countries, and inform individuals to reflect on our day-to-day media diet.

New Datasets

- WIKITRAFFIC [87], the largest dynamic network of time series with 366K nodes and 22M time-dependent links spanning five years. This dataset provides an open benchmark for developing models in this area, with applications that include optimizing resources for the web. Specifically, WIKITRAFFIC dataset is significantly larger than VEVOMUSIC, consisting of 61K nodes (music videos) and 300K+ links, released by our prior AFOSR work [96].

- COVID2020 [99], a new, longitudinal Twitter dataset during the first year of COVID-19, with more than 95% coverage using our own multi-stream data collection strategy [97], 17M+ geo-located users globally, and 346M+ tweets over an 8-month period.

Publication list

The following publications result from this project.

- [103] Haiqing Zhu, Yun Kuen Cheung, Lexing Xie. *Stability and Efficiency of Personalised Cultural Markets*. Under review at ACM TheWebConf, November 2022.
  Summary: We propose a novel view of attention allocation via computational economics

and mathematical optimization. Contrary to conventional wisdoms on the randomness of cultural markets, we show that a large number of users responding to popularity signals individually can lead to probabilistically stable and predictable outcomes that trades off market efficiency and diversity.
Technical sections: Section 1

- [87] Alasdair Tran, Alexander Patrick Mathews, Cheng Soon Ong and Lexing Xie. *Radflow: A Recurrent, Aggregated, and Decomposable Model for Networks of Time Series*. In *WWW '21: The Web Conference 2021*, Pages 730-742, 2021. https://doi.org/10.1145/3442381.3449945
  Summary: This work introduces a novel model *Radflow*, for effective and efficient forecasting and imputation on a network of time series. *Radflow* embodies three key ideas: a recurrent neural network to obtain node embeddings that depend on time, the aggregation of the flow of influence from neighboring nodes with multi-head attention, and the multi-layer decomposition of time series. We show that *Radflow* can learn different trends and seasonal patterns, that it is robust to missing nodes and edges, and that correlated temporal patterns among network neighbors reflect influence strength.
  Technical section: Section 2

- [99, 98] Cai Yang, Siqi Wu, Lexing Xie. *The Shapes of the Fourth Estate Worldwide: Profiling Global Media Audience During the COVID-19 Pandemic*. Under revision at ACM CSCW, October 2022.
  Summary: We curate a high-fedelity COVID dataset and study user-media interaction at a global scale. We map user political leaning on the top 20 countries by geo-located Twitter population, and present detailed statistical profiles of media reach and audience leaning for both the developed countries and the Global South.
  Technical section: Section 3

Technical Section

With information produced in abundance, human attention has become the scarce resource for which various types of information compete [83]. The internet operates as a major exchange of information and attention for the past few decades, yet surprisingly little is known about how individual choices and collective attention interact, or about how different parties can influence or control it. This important knowledge gap poses great challenges to the trustworthiness of online information from global political outcomes, to the mental well-being of individuals, to having reliable information about topics ranging from climate change to public health. The challenges are so great, that the study of collective attention is recently referred to as a "crisis discpline" [9]. This project aims to uncover the mathematical underpinnings of online attention market, design computational methods for making attention allocation stable and fair, enable applications where content producers, consumers, platforms and regulators could each play a role in influencing attention allocation towards socially beneficial goals. The new knowledge will enable organisations and individuals to foster safer and more trustworthy online spaces for the public, and benefit citizens through the minimisation of harm and misinformation when they are online.

The research program at Computational Media Lab of ANU has been advancing three frontiers in understanding online attention for the past 8+ years. The newest front, started during this project is a theoretical analysis of distributed behavior of social media users, and mapping the optimization landscape of well-known feedback loops in social networks (Section 1). In algorithm development, we develop new approaches to modeling networks of time series, surpassing state of the art models in both forecasting and imputation, and are able to scale to 100,000s of nodes, which no prior work has achieved (Section 2) – this work builds upon the machine learning and algorithmic work in our previous grants (`#1514018` and `#1914078`). In large-scale measurement, we map political leaning and media consumption at the global scale, providing a detailed statistical profile of media consumption and audience distribution for 20 countries spanning the global North and South (Section 3).

Full technical details are in our papers and pre-prints [87, 99, 103].

1 Theory: An optimization view of attention markets

In 2006 Salganik et al published seminal experiments on the effect of feedback signals on attention allocation among a set of songs, now dubbed MusicLab [79]. Several mathematical models have been proposed to describe it [67, 58], but all of them assumes that preferences to different items are the same for all users (via a term called *quality*), and none of them reveal the overall objective function of asynchrous behaviour from a large number of users. The level of mathematical understanding largely lags behind the the rich literature and wide-spread practice of recommender systems that estimates and caters to inividual user preferences. On the other hand, recent work in classical markets (especially Fisher markets) offer a range of results to understand equilibria from an algorithmic and optimization perspective [28, 100].

1.1 New insight: the objectives of MusicLab dynamics

We first define the simple *MusicLab* market. A set of items $\mathcal{I}$ is available for consumption, indexed by $i = 1, \ldots, |\mathcal{I}|$, each with a (fixed and known) quality factor $\bar{q}_i > 0$. The notation $\bar{q}_i$ is meant to emphasise that the notion of *average* quality across all users being modelled here (which is obviously simplistic). $\bar{q}_i$ is also an inseparable combination of *item ranking* and *inherent quality/preference* that can be modelled separately in a general market. An infinite number of users arrive at the market, and time t advances upon arrival of a new user. We use $\boldsymbol{\phi}^t \in [0,1]^{|\mathcal{I}|}$ to denote a vector of *market shares* of all items, where ϕ_i^t is the fraction of total attention that item i obtained up to time t. Naturally $\boldsymbol{\phi}^t$ lies on the interior and boundaries of a simplex, $\sum_i \phi_i \leq 1$, denoted as Δ. The probably that item i is chosen by the $(t+1)$-th user is descirbed by a multinomial logit model [**?** , Chapter 17]. Note that this probability is proportional to the item quality factor $\bar{q}_i$ and the current market share of the item ϕ_i^t warped by a constant exponent r, describing a feedback loop from the current market allocation. This mechanism is similar to what is known as *proportional response* in Fisher markets [100]. Then the market share of the next time step is simply $\boldsymbol{\phi}^{t+1} = g(\boldsymbol{\phi}^t) \in [0,1]^{|\mathcal{I}|}$, with:

$$g_i(\boldsymbol{\phi}^t) = \frac{\bar{q}_i (\phi_i^t)^r}{\sum_{k=1}^{|\mathcal{I}|} \bar{q}_k (\phi_k^t)^r}. \tag{1}$$

This choice process with feedback naturally leads to the notion of *market equilibrium* – a market share $\boldsymbol{\phi}$ is an equilibrium if $g(\boldsymbol{\phi}) = \boldsymbol{\phi}$.

The new insight is in establishing a robust connection between the equilibria of such a *homogeneous MusicLab* market and the following two constrained optimization problems.

Sum-utility

$$\max_{\boldsymbol{\phi} \in \Delta} \quad \sum_{i=1}^{|\mathcal{I}|} \bar{q}_i \phi_i^r \tag{2}$$

Regularised log-utility

$$\max_{\boldsymbol{\phi} \in \Delta} \quad \sum_{i=1}^{|\mathcal{I}|} [\phi_i \log \bar{q}_i - (1-r)\phi_i \log \phi_i] \tag{3}$$

One can view the *sum-utility* objective (2) as the "total utility" since the choice probability of item i is proportional to the "utility" all users obtain from it, which is $\bar{q}_i \phi_i^r$. When $0 < r \leq 1$ *sum-utility* is concave, leading to a unique maximum. The *regularised log-utility* objective (3)

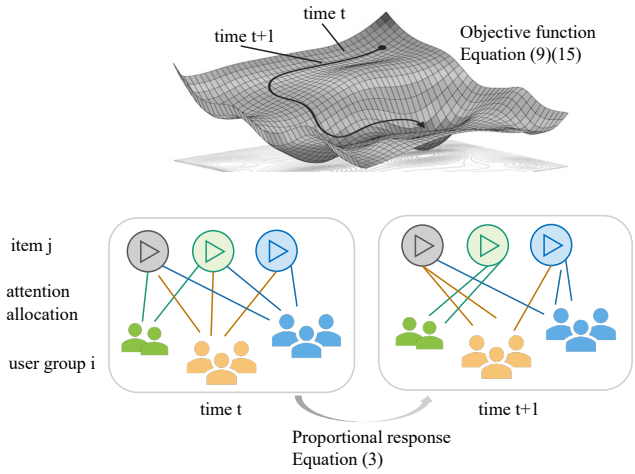


Figure 1: A core contribution of this work is to provide an optimisation view (Top) of cultural markets (Bottom), which affords new results on stability, efficiency and equilibrium behavior. (Bottom) An illustration of a culture market with several types of users interacting with a few items (color similarity between users and items indicate differing matches in preferences). Users allocate attention to the items based on proportional-response-esque dynamic.

has two parts. The first summation can be viewed as an alternative measure of total utility, with the item utility $\log \bar{q}_i$ weighted by its market share ϕ_i. This term encourages items with a larger quality term to have larger market shares. The second term is the information entropy of market share $H(\boldsymbol{\phi}) = -\sum_i \phi_i \log \phi_i$, encouraging $\boldsymbol{\phi}$ to be *flat* - i.e. spread among a wider set of items. Hyper-parameter r controls the regularisation strength between the two terms. When $0 < r \leq 1$ *regularised log-utility* is concave and therefore having a unique maximum, due to the first term being linear in $\boldsymbol{\phi}$, and entropy $H(\boldsymbol{\phi})$ being concave.

The equivalence of equilibria of (1) and the maximiser of *sum-utility* can be established by seeking the stationary point of the Lagrangian, leading to the solution $\phi_i^* = \frac{\bar{q}_i^\gamma}{\sum_j \bar{q}_j^\gamma}$, with exponent $\gamma = \frac{1}{1-r}$. One can also verify that the same $\boldsymbol{\phi}^*$ maximises *regularised log-utility* (by looking at derivatives of the Lagrangian). But *regularised log-utility* affords an additional benefit: $\boldsymbol{\phi}^{t+1} = g(\boldsymbol{\phi}^t)$ is the same as mirror descent update rule on this objective. The unpublished proof is inpired by a similar result on Fisher markets [28]. It additionally needs the theory of stochastic approximations and Robbins-Monro algorithm, and is omitted due to space constraints.

This finding may appear surprising, since the social network literature does not spell out objectives of distributed actions or the uniqueness of equilibria. More generally, the vast literature on new algorithms and empirical results on recommender systems and machine learning for social media is large disconnected from the theoretical work in markets and social network algorithms.

1.2 Our Contributions

The main themes of this work is to establish several robust connections between stochastic T-O markets and optimization, and to use these connections to rigorous show that the influence dynamics in these markets are stable. For the homogeneous markets, we discover two objective functions for which the equilibrium of a T-O market is maximizer of these objectives. The first objective

is the "total utility" of the market. The second objective is of particular interest due to its natural interpretation. It is a weighted sum of the efficiency and the diversity of the market shares in the market, as measured by the Shannon entropy. While efficiency is a natural benchmark, diversity in cultural market is also important for the healthy development of the platform. The diversity not only broaden the customer base, it also provides financial support to the less popular producers to keep them in the cultural industry. Thus, it is of the platform providers' interest to strike for a balance between efficiency and diversity.

Interestingly, we show that the influence dynamic is indeed equivalent to stochastic mirror descent on the second objective. This suggests the dynamic is implicitly optimizing a natural objective in the market. A significant consequence is this allows us to present a new proof of a result of Maldonado et al. [67] that the dynamic converges to the unique equilibrium of the market almost surely.

For heterogeneous markets, we show that the equilibrium is optimizing an *ex-post* version of Nash social welfare. In classical Fisher market setting, Nash social welfare is the product of users' utilities, whereas each utility is raised to a power of the user's budget. In our case, the power is the budget times the *efficiency* for that user at the equilibrium. Then we turn our focus to two interesting sub-classes of heterogeneous markets, namely (i) the users have the same appeals on the items, but they perceive the qualities of the items differently; (ii) the users perceive same qualities on the items, but they have different appeals on the items. For (i), we show that it is equivalent to a homogeneous market. For (ii) we design a new objective function, where the influence dynamic is equivalent to stochastic mirror descent on the objective. Again, this allows use to show the dynamic converges to a equilibrium almost surely.

The robust connection between the dynamics and optimization processes echoes with *the self-reinforced efficiency* of some economic systems, for which there exist natural dynamics or algorithms that can attain equilibrium, while the equilibrium optimizes popular efficiency measure like social welfare. See Related Work below for more relevant discussions.

We also empirically explore how ranking strategies deployed by platforms can affect the overall efficiency and diversity of the market via interactions of visibility and the underlying user preferences. Using the well-known MovieLens-100K dataset, we examine the (user-centric) efficiency and (item- or producer- centric) diversity measures across three ranking strategies: random, quality-driven, and popularity-driven. Results confirm quality ranking being superior among all for both metrics, while popularity ranking comes last. The seemingly slow convergence in the simulation points to open questions for analysing the rate in culture market optimisations.

1.3 Related Work

As early as in 1971, Simon [83] pointed out that in an information rich world, attention becomes the new scarcity that information consumes. Examples of attention economies include entertainment such as music, film and television [79, 12], political campaigns and votes [17], scientific publications and researchers [45]. Since Simon's visionary statement, the research community has formulated economic question about attention in a number of different ways. Such as articulating the phenomenon of attention scarcity in corporate life [36], diagnostic criteria for attention scarcity and solving it as (one off) allocation problems [42, 43], or connecting attention allocation to advertising revenue [41]. A recent study by Vosoughi et al. [90] showed that false news spreads faster online, suggesting that besides quality, appeal (e.g. novelty of the false news and the emotion it

stimulates) of a digital good is crucial in social influence. More broadly, the web research community have measured attention to items by individual users [85], a large set of users [95], and attention among a network of items [96].

The concept of self-reinforced efficiency of economic systems can be traced back to the "invisible hand" metaphor of Adam Smith. One of the first analytical confirmations of the concept is the famous First Welfare Theorem, which states that a market equilibrium of any complete market is Pareto efficient [6, 37]. Furthermore, in a broad class of markets called Eisenberg-Gale Fisher markets, market equilibrium optimizes a popular efficiency measure called Nash social welfare [40, 39, 51]. On the other hand, in combinatorial auction, any Walrasian equilibrium (if exists) optimizes the social welfare [15]. In many of these economic systems, there are natural adaptive price/bidding dynamics (e.g. tâtonnement [91, 7, 31, 32, 27, 26] and proportional response [92, 62, 100, 16, 28, 18, 46, 19, 29, 30]) or auction algorithms (e.g. ascending price auction [56, 73]) that can attain the efficient equilibria. As we shall see, the influence dynamics we study are indeed a stochastic version of proportional response.

1.4 Model: The Trial-Offer Market with Heterogeneous User Types

First, we describe stochastic trial-offer (T-O) market, in which users come to the platform one-by-one to try and purchase the items. This model generalises cultural markets specified by Krumme et al. [58] and Maldonado et al. [67] to heterogeneous types of users. Then we describe a continuous and deterministic analogue of the stochastic model which will be useful for analysis. In this work, we still use *purchase* to denote user completing a transaction on an item, where the resource a user *spends* is attention. One could think of it as a unit amount of time. Without loss of generality, we assume that each user has the same *budget* of attention, and that each item *costs* one unit of attention.

Stochastic Model. Let $\mathcal{U}$ denote the types of users and $\mathcal{I}$ denote the set of items. The fraction of Type-i users is denoted by w_i; note that $\sum_{i=1}^{N} w_i = 1$. If $|\mathcal{U}| = 1$, we say the market is *homogeneous*, otherwise it is *heterogeneous*.

The dynamic starts at time $t = 0$. At each time $t \geq 1$, a random user comes to the platform and tries an item, and then she decides to purchase the item or not. Let d_j^t denote the number of purchases of item j up to time t. To ensure that all items have a positive probability to be tried in the initial rounds, we assume that each item was purchased before the dynamic starts, i.e. $d_j^0 \geq 1$ for every item j. The *market share* of item j at time t is

$$\phi_j^t := \frac{d_j^t}{\sum_{k=1}^{|\mathcal{I}|} d_k^t}.$$

We let Δ denote the simplex of all possible market shares:

$$\Delta = \left\{ \boldsymbol{\phi} \in \mathbb{R}^{|\mathcal{I}|} : \sum_{j=1}^{|\mathcal{I}|} \phi_j = 1, \phi_j \geq 0 \text{ for any item } j \right\}. \tag{4}$$

If the user at round t is of Type-i, the probability that she will try item j is modelled as a

multinomial logit:

$$\frac{v_{ij}(\phi_j^{t-1})^{r_i}}{\sum_{k=1}^{|\mathcal{I}|} v_{ik}(\phi_k^{t-1})^{r_i}}\,, \tag{5}$$

where each $v_{ij} \geq 0$, and $r_i > 0$ is a parameter that depicts the strength of feedback signal for Type-i users.[1] v_{ij} is a parameter that depicts the *visibility* of item j to Type-i users, which depends on the *appeal* of the item itself, and also how the item is promoted or ranked with respect to other items. After a Type-i users tries an item j, she purchases the item with probability $q_{ij} \in [0, 1]$. This probability is intrinsic to the user and the item, and does not depend on exogenous factors.

In the homogeneous case, since there is only one user type, we will drop all indices i from the notations, resulting in v_j, q_j, r, and clearly $w_1 = 1$.

This distributed dynamic is the result of two interacting factors: A social feedback signal $(\phi_j^{t-1})_i^r$ based on the overall popularity of the item, such as the one implemented by the original musiclab experiment [79], or the number of downloads and likes on myriads of internet platforms; the user-specific visibility and quality factors v_{ij} and q_{ij} generalises recent models that analyse homogeneous attention market with feedback loop [58, 53], making it more realistic. This feedback dynamic is also similar to proprotional response in Fisher markets [28], which we will exploit for obtaining key results in Section 1.5.

Ranked list is one of the most popular forms of presenting a set of items to users, and a salient factor affecting the visibility of an item is its position in such a list [35]. If the positions are fixed throughout the attention dynamic, v_{ij} remains constant. Our theoretical results focus on this case. In our empirical simulations, we will explore how different strategies of positioning the items by the platform provider will affect the outcome.

We compute the probability of item j being the next purchase by manipulating the trial and purchase probabilities.

Lemma 1 *In a stochastic T-O market defined above, the probability that the next purchase is for item j, denoted by $p_j(\boldsymbol{\phi})$, is a function of the current market share $\boldsymbol{\phi}$, given by $y_j(\boldsymbol{\phi})/(\sum_{k=1}^{|\mathcal{I}|} y_k(\boldsymbol{\phi}))$, where*

$$y_j(\boldsymbol{\phi}) \;:=\; \sum_{i=1}^{|\mathcal{U}|} w_i q_{ij} \cdot \frac{v_{ij}(\phi_j)^{r_i}}{\sum_{k=1}^{|\mathcal{I}|} v_{ik}(\phi_k)^{r_i}}. \tag{6}$$

$y_j(\boldsymbol{\phi})$ represent the probability that item j is tried and then purchased by any user group. In the homogeneous case, the probability that the next purchase is for item j is

$$\frac{v_j q_j(\phi_j)^r}{\sum_{k=1}^{|\mathcal{I}|} v_k q_k(\phi_k)^r}. \tag{7}$$

For $\boldsymbol{\phi}$ to be a stationary point in this stochastic process, it must satisfy that $p_j(\boldsymbol{\phi}) = \phi_j$ for all items j. This motivates the following equilibrium notion.

Definition 2 *For any T-O market, we say a market share $\boldsymbol{\phi}$ is a trial-offer market equilibrium (TOME) if $\mathbf{p}(\boldsymbol{\phi}) = \boldsymbol{\phi}$.*

[1] Here $r_i = 0$ denotes no *social* feedback signal from the current market share, whereas $r_i \to \infty$ means only the most popular item will be chosen in the next round. If the denominator $\sum_{k=1}^{|\mathcal{I}|} v_{ik}(\phi_k^{t-1})^{r_i} = 0$, then this probability is defined as 0.

We show that TOME exists under mild conditions ([4]**??**) using the Brouwer's fixed-point theorem.

Theorem 3 *If (i)* $0 < r_i < 1$ *and* $w_i > 0$ *for any user type* i*, and (ii) for each item* j *there exists type* i *such that* $v_{ij}q_{ij} > 0$*, then the market must have a TOME* $\boldsymbol{\phi}^*$*, in which* $\phi^*_j > 0$ *for all items* j*.*

In the homogeneous case, we can explicitly compute the unique TOME $\boldsymbol{\phi}^*$ if $0 < r < 1$:

$$\phi^*_j \;=\; \frac{(v_j q_j)^{1/(1-r)}}{\sum_{k=1}^{|\mathcal{I}|}(v_k q_k)^{1/(1-r)}}, \tag{8}$$

provided that $v_k q_k > 0$ for some item k. It is easy to verify $\boldsymbol{\phi}^*$ is a stationary point by plugging it into Theorem 2, Section 1.5 specifies how to obtain $\boldsymbol{\phi}^*$ and argues for its uniqueness.

Efficiency and Diversity Measures. An online platform may be interested in maximizing $\sum_{j=1}^{|\mathcal{I}|} y_j(\boldsymbol{\phi})$, the probability of successful transaction among all items.

Definition 4 *At any market share* $\boldsymbol{\phi}$*, define the T-O market efficiency as* $E(\boldsymbol{\phi}) := \sum_{j=1}^{|\mathcal{I}|} y_j(\boldsymbol{\phi})$*.*

A platform may also be interested in promoting a diversity of items. A natural measure of diversity is the *Shannon entropy*, which is the standard measure of uncertainty of a probability distribution in information theory [33]. Given market share $\boldsymbol{\phi}$, its Shannon entropy is

$$H(\boldsymbol{\phi}) = -\sum_{j=1}^{|\mathcal{I}|} \phi_j \log \phi_j. \tag{9}$$

Continuous and Deterministic Analogue of the Stochastic Model. In this analogous model, there is one user of each type i, whose budget is w_i. The item set is again $\mathcal{I}$. The budget w_i corresponds to the maximum amount of attention the buyer can afford in the platform. At each time $t \geq 0$, each buyer i spends an amount of b^t_{ij} for item j. Subject to budget constraints, $\sum_{j=1}^{M} b^t_{ij} \leq w_i$. Let the total spending on each item be $b^t_j := \sum_{i=1}^{N} b^t_{ij}$, then the market share of item j at time t is $\phi^t_j := b^t_j/(\sum_{k=1}^{M} b^t_k)$. For $t \geq 1$, the update rule is

$$b^t_{ij} = w_i q_{ij} \cdot \frac{v_{ij}(\phi^{t-1}_j)^{r_i}}{\sum_{k=1}^{M} v_{ik}(\phi^{t-1}_k)^{r_i}} \;=\; w_i q_{ij} \cdot \frac{v_{ij}(b^{t-1}_j)^{r_i}}{\sum_{k=1}^{M} v_{ik}(b^{t-1}_k)^{r_i}}. \tag{10}$$

Note that the middle term in (10) is the same as the term appeared in (6), so $b_j = \sum_{i=1}^{|\mathcal{U}|} b_{ij}$ is the same as $y_j(\boldsymbol{\phi})$ in (6). With this, it is not hard to see that $\boldsymbol{\phi}^*$ is a TOME of a stochastic T-O market if and only if $\mathbf{b}^*$, where each b^*_{ij} is computed by the middle term of (10) by replacing $\boldsymbol{\phi}^{t-1}$ with $\boldsymbol{\phi}^*$, is a fixed point of the dynamic (10) of the analogous continuous and deterministic market.

Comparison with Classical Fisher Market and Proportional Response. The continuous and deterministic market and the dynamic (10) are reminiscent of the classical Fisher market and the proportional response (PR) dynamic in [28]. However, there is one crucial difference. PR dynamic in Fisher market is same as (10) but with b^{t-1}_k on the RHS replaced by b^{t-1}_{ik}/b^{t-1}_k. The term b^{t-1}_j is viewed as the *price* of item j, while a higher price in PR drives down spending on that item from the buyers. In contrast, a higher value of b^{t-1}_j in (10), which corresponds to receiving more attention in the T-O market, will lead to more spending on that item. This reflects the tendency of cascading in general influence dynamics.

1.5 Results

This section provides an overview of our key new results in four parts. Section 1.5.1 establishes a novel connection between TOME in homogeneous markets and two convex objectives. Section 1.5.2 shows that update steps in deterministic T-O markets are mirror descent steps for these objectives. Section 1.5.3 defines Robbins-Monro algorithm in order to connect the stochastic dynamics to the established optimisation framework. Section 1.5.4 presents the objective functions for heterogenous markets and its two special cases.

1.5.1 TOME maximises regularised utilities

First, we establish a robust connection between TOME of *homogeneous* T-O market and optimization. For notational simplicity, let $\bar{q}_j = q_j v_j$, noting that $q_j v_j$ are coupled in both (7) and (8). We consider the following constrained optimization problem:

$$\begin{aligned} \max \quad & \sum_{j=1}^{|\mathcal{I}|} \bar{q}_j \phi_j^r, \\ \text{subject to } & \boldsymbol{\phi} \in \Delta. \end{aligned} \tag{11}$$

We can view the objective function as the "total utility" since the choice probability of item j is proportional to the "utility" associated with it, which is $\bar{q}_j \phi_j^r$. We establish the equivalence between the equilibria and the maximisers of the above problem in the following theorem (proof in **??**).

Theorem 5 *If $\boldsymbol{\phi}^*$ is a maximiser of the problem (11), then it is a TOME.*

We also discover another optimization problem which captures the TOME in the homogeneous market. Consider the following constrained optimization problem:

$$\begin{aligned} \max \quad & \Psi(\boldsymbol{\phi}) = \sum_{j=1}^{|\mathcal{I}|} \left(\phi_j \log \bar{q}_j - (1-r)\phi_j \log \phi_j \right), \\ \text{subject to } & \boldsymbol{\phi} \in \Delta. \end{aligned} \tag{12}$$

The objective function in (12) can be decomposed into two parts, namely $\sum_{j=1}^{|\mathcal{I}|} \phi_j \cdot \log \bar{q}_j$ and $(1-r) \sum_{j=1}^{|\mathcal{I}|} -\phi_j \log \phi_j$. The first summation can be viewed an alternative measure of total utility, with the utility of item j being $\log \bar{q}_j$ weighted by its market share ϕ_j. The second is $(1-r)H(\boldsymbol{\phi})$ – the entropy of market share. When $r = 1$, the entropy term disappears, so the optimization problem (12) becomes trivial: the optimal solution is by setting $\phi_j = 1$ for the highest-utility item $j = \arg\max_k \bar{q}_k$. As r decreases from 1, i.e. the strength of feedback signal reduces, the entropy term becomes more significant, which encourages diversity in the optimal solution.

For $0 < r < 1$, (11) and (12) are both concave in $\boldsymbol{\phi}$, therefore having a unique maximum. A crucial advantage of (12) over (11) is that mirror descent on (12) provides insight into the convergence of the stochastic influence dynamics in T-O market as specified in (5) and (7). To formally describe this discovery, we need several concepts in optimization theory, which are discussed next.

1.5.2 T-O update as mirror descent

Consider a general constrained convex optimization problem of minimizing a smooth convex function $f(x)$, subject to the constraint $x \in C$ for some convex set C.

Definition 6 *Let C be a compact and convex set. Let h be a differentiable convex function on C. The* Bregman divergence *w.r.t. h, denoted by d_h, is defined as*

$$d_h(x, y) = h(x) - h(y) - \langle \nabla h(y) \ , \ x - y \rangle,$$

for any $x \in C$ and $y \in \mathit{rint}(C)$.

The famous Kullback–Leibler (KL) divergence is a special case of Bregman divergence, generated by the function $h(x) = \sum_j (x_j \log x_j - x_j)$.

Given a Bregman divergence d_h, the corresponding mirror descent update rule is

$$x^t \ = \ \arg\min_{x \in C} \left\{ \langle \nabla f(x^{t-1}) \ , \ x - x^{t-1} \rangle + \frac{1}{\alpha} \cdot d_h(x, x^{t-1}) \right\}, \tag{13}$$

where α is considered as the step-size of the update rule, which may depend on t in general.

In the next section, we will discuss conditions on f for which (13) converges to the optimal solution. But before that, we point out the following crucial discovery.

Lemma 7 *Let $\boldsymbol{\phi}^t$ be market share over time, and function $\mathbf{p}(\boldsymbol{\phi})$ as defined in Theorem 1. The update rule*

$$\boldsymbol{\phi}^t = \mathbf{p}(\boldsymbol{\phi}^{t-1}) \tag{14}$$

is same as the mirror descent update rule (13) for the optimization problem (12), in which d_h to taken as the KL divergence, and $\alpha = 1$.

The update rule (14) is deterministic, but we can use the methodology in Maldonado et al. [67] to rewrite the stochastic influence dynamic as a Robbins-Monro algorithm. Then we can use results in stochastic approximation [13] to establish convergence of the influence dynamic. Proof for this lemma is in the supplemental material [4].

1.5.3 Robbins-Monro Algorithm

Definition 8 *([13, 77]) A Robbins-Monro algorithm (RMA) is a discrete-time stochastic process z^t whose general structure is specified by*

$$z^t - z^{t-1} \ = \ \gamma^t \cdot \left(F(z^{t-1}) + U^t \right),$$

where $z^t \in \mathbb{R}^n$ for some $n \geq 1$, $F : \mathbb{R}^n \to \mathbb{R}^n$ is a deterministic continuous vector field, γ^t is deterministic and satisfies $\gamma^t > 0$, $\sum_{t \geq 1} \gamma^t = +\infty$ and $\lim_{t \to \infty} \gamma^t = 0$, and $\mathbb{E}[U^t | \mathcal{F}^{t-1}] = 0$ where $\mathcal{F}^{t-1}$ is the natural filtration on the entire process.

Note that market share $\boldsymbol{\phi}$ will change only when there is a purchase. Thus, Maldonado et al. [67] modify the time schedule to only count those times at which a purchase occur, and show the following lemma.

Lemma 9 ([67]) *In the stochastic T-O market, the update of market share follows the following RMA w.r.t. the modified time schedule:*

$$\boldsymbol{\phi}^t - \boldsymbol{\phi}^{t-1} \;=\; \frac{1}{t} \cdot \left((\mathbf{p}(\boldsymbol{\phi}^{t-1}) - \boldsymbol{\phi}^{t-1}) + U^t \right),$$

where U^t is the random variable defined as below. Let $\mathbf{e}^t$ denote the random unit vector whose j-th entry is 1 if item j is purchased at time t. Then $U^t = \mathbf{e}^t - \mathbb{E}[\mathbf{e}^t | \mathcal{F}^{k-1}]$. (Recall that $e_j^t = 1$ with probability $p_j(\boldsymbol{\phi}^{t-1})$.)

With the above lemma, we can apply the seminal results of Benaïm [13] to show that the RMA trajectory is the asymptotic pseudo trajectory of the mirror descent update (14). By using the convergence theorem established in [25, 28], we show that both dynamics converge to the global minimisers of (12). This allows us to present a new proof of the following theorem in [67] (see [4]**??** for details).

Theorem 10 ([67], Theorem 5.3) *With the homogeneous setup, if $\boldsymbol{\phi}^0 > 0$ and $0 \leq r < 1$, then with probability 1,*

$$\lim_{t \to \infty} \boldsymbol{\phi}^t = \boldsymbol{\phi}^*, \tag{15}$$

where $\boldsymbol{\phi}^*$ is uniquely defined as the maximiser of the convex program (12). And for any $r \in [0, 1]$,

$$\lim_{t \to \infty} \Psi(\boldsymbol{\phi}^t) = \Psi^* \tag{16}$$

with probability 1, where Ψ is defined in (12), and Ψ^* is the global maximum of problem (12).

1.5.4 TOME for heterogeneous markets

For the heterogeneous case, we show the following proposition, which depicts that the TOME is optimizing an ex-post version of a convex objective.

Proposition 11 *The optimal solution of the following convex program is the TOME:*

$$\begin{aligned} \max_{\{\phi_j\}_{j=1}^{|\mathcal{I}|}} \prod_{i=1}^{|\mathcal{U}|} \Big(\sum_{j=1}^{|\mathcal{I}|} q_{ij} v_{ij} \phi_j^r \Big)^{w_i a_i^*}, \\ \textit{subject to:} \quad \sum_{j=1}^{|\mathcal{I}|} \phi_j = 1 \end{aligned} \tag{17}$$

where $a_i^* = \frac{\sum_{j=1}^{|\mathcal{I}|} q_{ij} v_{ij} \phi_j^*}{\sum_{k=1}^{|\mathcal{I}|} v_{ik} \phi_k^*}$ and ϕ_j^* satisfies the equilibrium equation for any $i \in \mathcal{U}$ and $j \in \mathcal{I}$.

A proof is provided in the online supplement [4]**??**. (11) takes the form of a product-of-utilities (or sum-of-log-utilities). Known as Nash social welfare [55], this utility affords fairness properties in the resulting allocations [14, 23]. A proper exposition of this connection is outside the scope of this work. Once a_i^* are known, (11) is convex in $\boldsymbol{\phi}$ – hence ex-post. In general, optimization involving a_i^* is non-convex, and we deem properties of its solution open problems.

Then we turn our focus to two interesting cases of heterogeneous market, where $r_i = r$ for all types i, and:

- the trial randomness is the same across all user types (i.e., v_{ij} are the same for all types i), but the purchase randomness can be different (i.e., q_{ij} can be different for various types i);
- the trial randomness can be different across all types (i.e., v_{ij} can be different for various types i), but the purchase randomness is the same across all types (i.e., q_{ij} are the same for all types i).

For the first case, we show that it can be reduced to the homogeneous case. For the second case, we design a new optimization problem and show that (10) is indeed mirror descent for the problem. The driving variables of the problem are b_{ij} but not ϕ_j, where b_{ij} are the driving variables in the dynamic (10). We let $b_j = \sum_{i=1}^{|\mathcal{U}|} b_{ij}$, and set q_j to be the common value of q_{ij} for all i. The optimization problem is

$$\begin{aligned} \max \quad & \sum_{j=1}^{|\mathcal{I}|} -\frac{r}{q_j}(b_j \log b_j - b_j) \\ & + \sum_{i=1}^{|\mathcal{U}|}\sum_{j=1}^{|\mathcal{I}|} \left(\frac{1}{q_j}(b_{ij}\log b_{ij} - b_{ij}) - \frac{b_{ij}}{q_j}\log q_j v_{ij} \right) \\ \text{subject to} \quad & \sum_{j=1}^{|\mathcal{I}|} \frac{b_{ij}}{q_j} = w_i, \ \forall i \in \mathcal{U}. \end{aligned} \tag{18}$$

By perform a variable transformation $x_{ij} = b_{ij}/q_j$ to the above problem, we obtain an equivalent transformed optimization problem where x_{ij} are the driving variables, which is needed for the key lemma below. The proof is in [4]**??**.

Lemma 12 *The update rule* (10) *is equivalent to mirror descent w.r.t. KL divergence on the transformed optimization problem.*

With the above lemma in hand, we again use RMA and Benaïm's results to establish convergence of the stochastic influence dynamics in heterogeneous market. The proof can be found in supplement [4]**??**.

Theorem 13 *With the heterogeneous setup when $r_i = r < 1$ for all $i \in \mathcal{U}$, if one of the following conditions*

1. *$v_{ij} = v_j$ for all $i \in \mathcal{U}, j \in \mathcal{I}$.*
2. *$q_{ij} = q_j$ for all $i \in \mathcal{U}, j \in \mathcal{I}$.*

is satisfied. In addition, $\boldsymbol{\phi}^0 > 0$. Then with probability 1,

$$\lim_{t\to\infty} \boldsymbol{\phi}^t = \boldsymbol{\phi}^*, \tag{19}$$

where $\boldsymbol{\phi}^*$ is uniquely defined as the maximiser of the convex program (18)*. And when $r \in [0, 1]$,*

$$\lim_{t\to\infty} \Gamma(\boldsymbol{\phi}^t) = \Gamma^*, \tag{20}$$

with probability 1, where Γ is the objective function of problem (18)*, and Γ^* is the global maximum of problem* (18)*.*

1.6 Analysis

The strategy to prove our main result Theorem 13 could be divided into two main steps. First, after applying the variable transformation $x_{ij} = b_{ij}/q_j$ mentioned in the last section, we demonstrate the equivalence between the stochastic RMA process and the deterministic "weighted" mirror descent dynamic. Then, we exploit the convexity of the proposed objective function to deduce the convergence.

1.6.1 RMA and asymptotic pseudotrajectory

First, we need the notion of asymptotic pseudotrajectories. It formulation needs us to introduce the concept of semiflow, which is a flow but only considering the part where $t \geq 0$.

Definition 14 ([13], section 3) *A semiflow Φ on a metric space (M, d) is a continuous map*

$$\begin{aligned} \Phi : \mathbb{R}_+ \times M &\to M, \\ \Phi(t, x) &= \Phi_t(x) \end{aligned} \tag{21}$$

such that

$$\Phi_0 = id, \ \Phi_{t+s} = \Phi_t \circ \Phi_s \tag{22}$$

for all $(t, s) \in \mathbb{R}_+ \times \mathbb{R}_+$.

Next, for a (semi)flow (which is potentially induced by a vector field), we define the notion of asymptotic pseudotrajactory. Roughly, it means that a trajectory is very close to the integral curve (or flow) if we push t to infinity.

Definition 15 ([13], section 3) *A continuous function $X : \mathbb{R}_+ \to M$ is an asymptotic pseudotrajectory for ϕ if for any $T > 0$,*

$$\lim_{t\to\infty} \sup_{0\leq h\leq T} d(X(t+h), \Phi_h(X(t))) = 0. \tag{23}$$

Now we turn back to our problem.. Our goal is to study the trajectory of a discrete-time process $\{z_n\}_{n\in\mathbb{N}} \subset \mathbb{R}^m$ adapted to the filtration $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ which could be written as

$$z_{n+1} - z_n = \gamma_{n+1}(F(z_n) + U_{n+1}), \tag{24}$$

where $F \in C^\infty(\mathbb{R}^m, \mathbb{R}^m)$ is a smooth map, $\{\gamma_n\}_{n\in\mathbb{N}}$ are the step sizes satisfying

$$\sum_{n=1}^{\infty} \gamma_n = \infty \quad \text{and} \quad \lim_{n\to\infty} \gamma_n = 0, \tag{25}$$

and $U_n \in \mathbb{R}^m$ are random perturbations satisfying $\mathbb{E}[U_{k+1}|\mathcal{F}_k] = 0$. For such a process, the "trajectory" could be defined as the interpolated curve connecting the sequence in $\mathbb{R}^m$ generated by the stochastic process. Set

$$\tau_0 = 0, \tau_n = \sum_{i=1}^{n} \gamma_i, \tag{26}$$

where $n \geq 1$. And define the continuous time affine interpolated process $Z(t)$ as

$$Z(\tau_n + s) = z_n + s\frac{z_{n+1} - z_n}{\tau_{n+1} - \tau_n} \tag{27}$$

for all $n \in \mathbb{N}$ and $0 \leq s < \gamma_{n+1}$. Finally, we have the main tool of analysing such the Robbin-Monro Algorithms, which is summarised in the following theorem,

Theorem 16 ([13], section 4) *Let F be a smooth vector field on M. In addition, for any point p it has unique integral curves around p. Assume that*

1. $\sup_n \mathbb{E}[\|U_{n+1}\|^q] \leq \infty$ *and* $\sum_{n\in\mathbb{N}} \gamma_n^{1+\frac{q}{2}} \leq \infty$ *for some* $q \geq 2$.
2. *Either*

$$\sup_n \|z_n\| \leq \infty \tag{28}$$

or F is Lipschitz and bounded on a neighbourhood of $\{x_n : n \geq 0\}$.

Then the interpolated process Z is an asymptotic peseudotrajectory of the flow Φ induced by F almost surely.

With these results, establishing the equivalence between the stochastic RMA and the deterministic weighted mirror descent can be done (see [4] **??** for details).

1.6.2 Convergence of mirror descent dynamics

As discussed in the preceding section, the key steps to demonstrate the convergence result rely on the strong convexity of our objective function for the mirror descent update. In particular, we will focus on the class of functions satisfying the following definition.

Definition 17 *The function f is (σ, L)-Bregman strongly convex with respect to the Bregman divergence d_h if for any $\mathbf{y} \in rint(C)$ and $\mathbf{x} \in C$,*

$$\begin{aligned} f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle + \sigma d_h(\mathbf{x}, \mathbf{y}) &\leq f(\mathbf{x}) \\ &\leq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle + L d_h(\mathbf{x}, \mathbf{y}). \end{aligned} \tag{29}$$

Thankfully, the objective functions of mirror descent updates for both homogeneous and heterogeneous cases are Bregman strongly convex. The proof can be found in [4]**??**.

Lemma 18 *With the heterogeneous setting when $r_i = r$ for all $i \in \mathcal{U}$, the objective function of problem (18) is $(1 - r, 1)$-Bregman strongly convex w.r.t the KL divergence.*

Next, to demonstrate the continuous analog (10) is converging, we could apply the following theorems for both the case of $r < 1$ and the case when $r = 1$.

Theorem 19 ([28], Theorem 3.1) *Suppose that f is (σ, L)-strongly Bregman convex with respect to d_h. With the update rule (13) together with the fact that $\alpha = \frac{1}{L}$, for all $t \geq 1$,*

$$f(\mathbf{x}^t) - f(\mathbf{x}^*) \leq \frac{\sigma}{(\frac{L}{L-\sigma})^t - 1} \cdot d_h(\mathbf{x}^*, \mathbf{x}^0). \tag{30}$$

Theorem 20 ([28], Theorem 3.2) *Suppose f is L-Bregman convex function with respect to d_h, and $\mathbf{x}^T$ is the point reached after T applications of the mirror descent update rule (13) together with the fact that $\alpha = \frac{1}{L}$. Then,*

$$f(\mathbf{x}^T) - f(\mathbf{x}^*) \leq \frac{L \cdot d_h(\mathbf{x}^*, \mathbf{x}^0)}{T}. \tag{31}$$

To prove the market share $\boldsymbol{\phi}^t$ converges as $t \to \infty$, we exploit the RMA theory and converted the problem to analysing the following "weighted" version of the mirror descent dynamic.

$$\mathbf{x}^{t+1}_{\text{weighted}} = \beta_t g(\mathbf{x}^t) + (1 - \beta_t)\mathbf{x}^t. \tag{32}$$

where $g(\mathbf{x}^t)$ is the standard mirror descent update satisfying $g(\mathbf{x}^t) = \arg\min_{x \in C} \left\{ \langle \nabla f(x^t)\ ,\ x - x^t \rangle + \frac{1}{\alpha} \cdot d_h(x, x$ $\{\beta_t\}_{t \in \mathbb{N}}$ are step sizes.

Beyond existing results, the convergence of update rule (32) is established in the following theorem (proof is provided in [4]**??**)

Theorem 21 *Suppose that f is (σ, L)-strongly Bregman convex ($\sigma, L \geq 0$) with respect to d_h. In addition, $d_h(\mathbf{x}, \mathbf{y})$ is convex on both variables. Under the conditions $\sum_{t=0}^{\infty} \beta_t = \infty$, and $0 < \beta_t \leq 1$ for any $t \geq 0, t \in \mathbb{Z}$, with the update rule (32), the dynamic converges to the global minimum of f, which is*

$$\lim_{t \to \infty} f(\mathbf{x}^t_{weighted}) = f(\mathbf{x}^*). \tag{33}$$

where the step size $\alpha_t = 1$.

1.7 Empirical observations

We simulate cultural markets using real-world preferences from the well-known MovieLens dataset [49], in order to explore the efficiency and diversity of the market with different ranking strategies under currently available information, such as by popularity [79] or quality.

We set up the simulation using the MovieLens-100k dataset [49] and construct a market with heterogeneous user types using two data sources. This dataset consists of 100,000 ratings (1-5) from 943 users on 1682 movies, where each user has rated at least 20 movies. In addition, simple demographic information such as gender and age for each user is provided. We performed matrix completion using incomplete SVD [12] via the `Surprise` python package[2], yielding a preference matrix $\Gamma = [\gamma_{ij}] \in \mathbb{R}^{943 \times 1682}$ for each (user, movie) pair. With $\gamma_{ij} \in [0, 1]$ denoting the normalised preference of user $i \in \mathcal{U}$ and item (movie) $j \in \mathcal{I}$.

To simulate heterogeneous perference types, we then divide the users into M non-overlapping subgroups based on user attributes. Let $\bigcup_{i=1}^{M} \mathcal{U}_i = \mathcal{U}$, where $\mathcal{U}_i$ denote the set of users in group i. Denote $\mathcal{O}$ as the set consists of all indices (i, j) with $\gamma_{i,j}$ observed (in the MovieLens-100k dataset), and $\bar{\mathcal{O}}$ as the set of unobserved indices (the remaining entries estimated with incomplete SVD). We first calculate the *internal visibility* factor $\hat{v}_{ij}$ by averaging over the set of observed entries generated by user group i using (34). This choice reflects the intuition that platforms rank items based on known user ratings. We also calculate the *quality* factor q_{ij} as by averaging over the set of *unobserved* entries generated by user group i using (35). This choice reflects the intuition that in a T-O market, the probability of purchase depend on a quality factor that is often visible before

[2] https://surpriselib.com

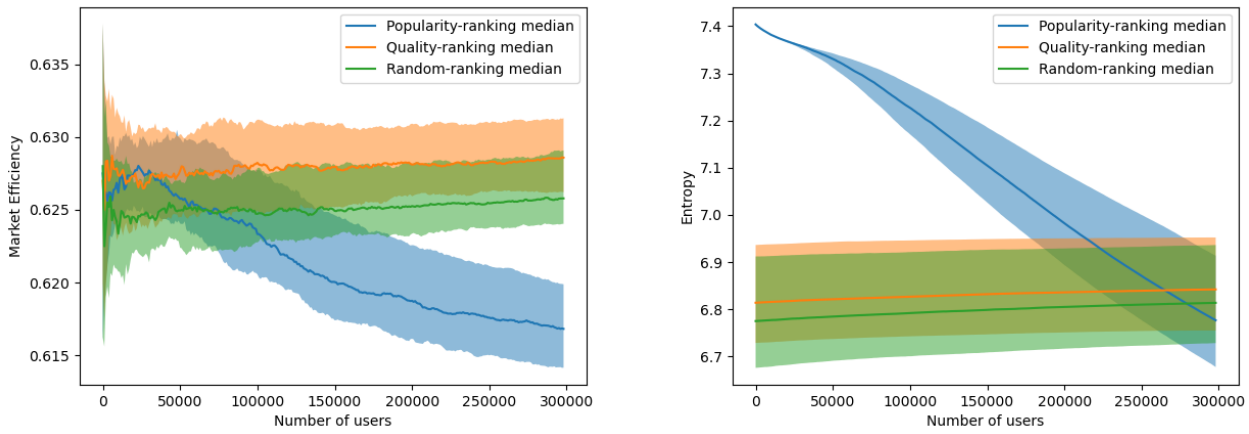


Figure 2: The evolution of market efficiency (left) and entropy (right) under different ranking strategies. Each simulation is run for 300,000 new users, a measurement is taken after every 1000 time steps. The solid lines denote the median, the shaded area represent the 25-th to the 75-th percentile of 50 repeated simulations.

a user tries an item. Averaging over unobserved entries makes q_{ij} contain the maximum amount of new information against $\hat{v}_{ij}$ within MovieLens, yet also ensures that q_{ij} and $\hat{v}_{ij}$ are positively correlated (rather than being negatively correlated, which can drive efficiency to zero).

$$\hat{v}_{ij} = \frac{1}{|(k,j) \in \mathcal{O}, k \in \mathcal{U}_i|} \sum_{(k,j)\in\mathcal{O}, k\in\mathcal{U}_i} \gamma_{kj}. \tag{34}$$

$$q_{ij} = \frac{1}{|(k,j) \in \bar{\mathcal{O}}, k \in \mathcal{U}_i|} \sum_{(k,j)\in\bar{\mathcal{O}}, k\in\mathcal{U}_i} \gamma_{kj} \tag{35}$$

Note that users in the simulation are *generalisations* of these M groups of user populations from MovieLens via attributes q_{ij} and $\hat{v}_{ij}$, rather than being subsets (or samples) from the original 943 users. Other strategies for estimating q_{ij} and $\hat{v}_{ij}$ is left for future work. In this section, we report results where users form two groups, splitting based on gender (only binary gender attribute is available in MovieLens). We use two groups to illustrate the differences in efficiency and diversity, results for other group compositions are in the supplement [4].

We also account for position bias in ranked lists [35] in order to construct the actual visibility factor v_{ij}, which can vary over time, denoted as v_{ij}^t. We use the position bias parameters from MUSICLAB model [58, 67], linearly interpolated to the size of MovieLens-100K. This results in a list of fixed weights ι_k, with $k = 1, \ldots, 1682$. We adopt the separable click-through rate (CTR) model commonly used in modeling auctions [1], which simply multiplies the inherent visibility with η_{ij}^t, the position bias factor of item j presented to user i at time t.

$$v_{ij}^t = \hat{v}_{ij}\eta_{ij}^t \tag{36}$$

The relationship between ι_k and η_{ij}^t is defined in the following three ranking strategies.

- Random-ranking. Upon each simulation round, the ranking of items is randomly and uniformly drawn from a permutation of indexes $\{1, 2, \ldots, 1682\}$. That is, the visibility term v_{ij} is set to the

product of $\hat{v}_{ij}$ and a random element of ι. This ranking changes for each simulation step. One expects such a strategy to promote diversity while preserving some information on item appeal through $\hat{v}_{ij}$.

- Popularity-ranking. This strategy sorts the items by descending market share $\boldsymbol{\phi}^t$, and set $\eta_{ij}^t = \iota_k$, with $k = \arg\text{sort}_{desc}\{\boldsymbol{\phi}^t\}[j]$. This ranking will change over the simulation steps, and is analogous to the original MUSICLAB experimental setting [79]. One expects this strategy to be unstable due to the randomness early in the simulation, since it could accidentally promote items that users do not like to the top resulting in the high quality items being buried.

- Quality-ranking. Denote the descending sorting rank of item j among user group i as $k = \arg\text{sort}_{desc}\{q_{i,:}\}[j]$, where $q_{i,:}$ is the one-dimensional array for qualities factors in group i, and $[j]$ denotes array indexing. Then the position bias factor is set to the corresponding item, and $\eta_{ij}^t = \iota_k$. This ranking does not change over the simulation steps, since both q_{ij} and its sorting order remains fixed. One expects this strategy to best align visibility with the underlying quality metrics (unobserved before trying), since it has *oracle* access to q_{ij}, and should yield high efficiency.

In the simulation, an infinite number of new users arrives over different rounds, some will make a purchase, some will try an item and *not* make a purchase (after flipping a biased coin parameterized by q_{ij}). This setting is consistent with other theoretical and simulation studies of cultural market and recommender systems [58, 53]. We report *market efficiency* as the empirical version of Theorem 4, namely, the fraction of users who made a purchase. We also measure diversity among the set of items, by computing entropy (9) at each time step.

Figure 2 summarises the trends of efficiency and diversity over the three ranking settings. The quality-ranking oracle has the highest efficiency and highest diversity overall, followed by random ranking, and popularity ranking has the lowest efficiency and diversity when there are sufficient number of users. For popularity ranking, we observe less than 100 items with any attention at all at the end of most simulation runs. This confirms our intuition that despite the presence of user-specific preference data, popularity-based rankings are unstable: prone to be driven by random initialisations, and result in market dominance by a few popular items. Note that the differences in empirical measures of efficiency is fairly small (within 0.02 of each other), implying that the relative importance of preference data and social signals can be quantified, which is left as future work. The fairly wide and over-lapping inter-quatile ranges of the three rankings could imply a slow convergence rate and high noise at 300,000 simulation steps, or it could be caused by the fully general objective function (17) being ex-ante non-convex and potentially not stable. We found it surprising *how slowly* the diversity of items by popularity decreased, it took more than 250,000 users for the diversity to drop below random or quality- ranking strategy for most simulation runs. This indicates that more understanding is needed for the convergence rate of stochastic heterogeneous T-O markets.

2 Algorithm: Modeling networks of time series

We propose a novel neural model for networks of time series that tackles all three challenges. We adopt a **R**ecurrent structure that affords time-sensitive **A**ggregations of network flow on top of the **D**ecomposition principle of time series; hence the name **Radflow**. It is more expressive

than N-BEATS [74] because it can generate node embeddings to handle graph inputs. It is more scalable than T-GCN as it can process hundreds of thousands of nodes via network attention and importance-weighted node sampling.

The structure of *Radflow* allows it to take dynamically changing nodes and edges as inputs, makes it tolerant to missing data, and is suitable for multivariate time series. Moreover, its multi-head attention strategy and layered decompositions provide interpretations over network influences and time.

In the following sections, we first describe the problem and then we introduce *Radflow* in detail.

2.1 Problem Statement

Consider the problem of time series forecasting in a graph. The input is a graph $G = (\boldsymbol{V}, \boldsymbol{E})$, consisting of N nodes denoted $\boldsymbol{V} = \{\boldsymbol{v}^1, \boldsymbol{v}^2, \ldots, \boldsymbol{v}^N\}$, and M edges. Each node $\boldsymbol{v}^j$ is associated with a multivariate time series having T observations:

$$\boldsymbol{v}^j = [\boldsymbol{v}_1^j, \boldsymbol{v}_2^j, \ldots, \boldsymbol{v}_T^j] \tag{37}$$

where $\boldsymbol{v}_t^j \in \mathbf{R}^D$ is the D-dimensional observation vector of node $\boldsymbol{v}^j$ at time step t. When the time series has only one value per time step (univariate), then $D = 1$. We use $\boldsymbol{v}_{[t:s]}^j$ to denote a subsequence of $\boldsymbol{v}^j$ containing all observations from time t to s, where $t \leq s$:

$$\boldsymbol{v}_{[t:s]}^j = [\boldsymbol{v}_t^j, \boldsymbol{v}_{t+1}^j, \ldots, \boldsymbol{v}_{s-1}^j, \boldsymbol{v}_s^j] \tag{38}$$

If node $\boldsymbol{v}^i$ has the potential to directly influence the time series of node $\boldsymbol{v}^j$ at time step t, then we add a directed edge e_t^{ij} from $\boldsymbol{v}^i$ to $\boldsymbol{v}^j$, and $\boldsymbol{v}^i$ becomes a neighbor of $\boldsymbol{v}^j$. We define $\mathbf{N}_t(\boldsymbol{v}^j)$ to be the set of neighbors of $\boldsymbol{v}^j$ at times step t. Edges may appear and disappear over time, thus G is a dynamic graph. We can now represent G as an adjacency array $\boldsymbol{A} \in \mathbf{R}^{N \times N \times T}$. For an unweighted directed graph, the entry a_{ijt} in $\boldsymbol{A}$ is 1 if edge e_t^{ij} exists, and 0 otherwise.

We now define the time series forecasting problem as it applies to dynamic graphs. The forecast length F is the number of future time steps for which the model will make predictions, while the backcast length B is the number of past observations available for making such predictions. Suppose we are currently at time $t = 0$. To forecast the time series of node $\boldsymbol{v}^j$ (which we shall call the *ego* node) from time step 1 to F, the forecast model

$$\hat{\boldsymbol{v}}_{[1,F]}^j = \text{ForecastModel}\left(\boldsymbol{v}_{[-B+1:0]}^j, \mathcal{V}^{\mathbf{N}(\boldsymbol{v}^j)}\right) \tag{39}$$

will take two sets of inputs: the B most recent observations of $\boldsymbol{v}^j$ and the information from its neighbors. This leads to two different settings, both of which will be evaluated in Section 1.5. The first is IMPUTATION, in which we observe the true values of the neighbors at the time of prediction. This amounts to using the ground-truth observations of the neighbors during the forecast period:

$$\mathcal{V}^{\mathbf{N}(\boldsymbol{v}^j)} = \left\{\boldsymbol{v}'_{[-B+1:F]} \;\middle|\; \boldsymbol{v}' \in \mathbf{N}(\boldsymbol{v}^j)\right\} \tag{40}$$

This is the setting used by [96] and is most useful when the main goal is to fill in missing data in the time series, or to interpret the influence between nodes. The second setting is FORECAST, where

we first use our best pure time series model to predict the future observations of each neighbor. These predictions are then used in the full model to forecast $\boldsymbol{v}^j$ itself.

In both settings, the final output of the model is

$$\hat{\boldsymbol{v}}^j_{[1,F]} = [\hat{\boldsymbol{v}}^j_1, \hat{\boldsymbol{v}}^j_2, \ldots, \hat{\boldsymbol{v}}^j_F] \tag{41}$$

corresponding to the forecast values for the next F time steps. Here we use the hat notation to denote the model's predictions, e.g., $\hat{\boldsymbol{v}}^j_t$ is the forecast time series vector of the ground truth $\boldsymbol{v}^j_t$.

2.2 Radflow

Radflow consists of two main modules: a recurrent component and a flow aggregation component. The recurrent component models all the time series in the graph independently, while the flow aggregation component additively adjusts the predictions based on the neighboring time series. The forecast $\hat{\boldsymbol{v}}^j_t$ of node $\boldsymbol{v}^j$ at time step t is obtained by summing up the outputs of the two main modules,

$$\hat{\boldsymbol{v}}^j_t = \hat{\boldsymbol{v}}^{jR}_t + \hat{\boldsymbol{v}}^{jA}_t \tag{42}$$

where $\hat{\boldsymbol{v}}^{jR}_t$ is the forecast contribution from the recurrent component and $\hat{\boldsymbol{v}}^{jA}_t$ is the contribution from the aggregation component. Note that $\hat{\boldsymbol{v}}^{jA}_t$ is itself a function of $\hat{\boldsymbol{v}}^{jR}_t$.

2.2.1 Recurrent component

We predict time series by breaking them down into L components using stacked recurrent blocks. The recurrent component is also designed to feed into the flow aggregation component which uses the node vectors to aggregate information in a neighborhood.

We first project the historical observations of the time series into a latent space in $\mathbf{R}^H$, where H is the hidden state size:

$$\boldsymbol{z}^{j1}_t = \boldsymbol{W}^D \boldsymbol{v}^j_t \tag{43}$$

Here $\boldsymbol{W}^D \in \mathbf{R}^{H \times D}$ is a learnable weight matrix. To get an intuitive justification for this projection, consider the special case where $D = 1$ and $\boldsymbol{W}^D$ is the all-ones vector. Then $\boldsymbol{z}^{j1}_t$ would contain H copies of the observation v^j_t. This resembles running an ensemble of H different time series models in parallel.

The recurrent component of our model consists of L blocks. Let $\boldsymbol{z}^{j\ell}_t$ be the input to Block$_\ell$ for node $\boldsymbol{v}^j$ at step t. In particular, the vector $\boldsymbol{z}^{j1}_t$ computed in Equation (43) will be used as the input to the first block. Each block will output three vectors—the backcast vector $\boldsymbol{p}^{j\ell}_t$, the forecast vector $\boldsymbol{q}^{j\ell}_t$, and the node vector $\boldsymbol{u}^{j\ell}_t$:

$$(\boldsymbol{p}^{j\ell}_t, \boldsymbol{q}^{j\ell}_t, \boldsymbol{u}^{j\ell}_t) = \text{Block}_\ell(\boldsymbol{z}^{j\ell}_t) \tag{44}$$

with $\boldsymbol{p}^{j\ell}_t, \boldsymbol{q}^{j\ell}_t, \boldsymbol{u}^{j\ell}_t \in \mathbf{R}^H$. Specifically inside each block, we have an LSTM cell followed by feedforward layers. The LSTM cell is first to operate, accepting as input: the time series residual computed by the previous block $\boldsymbol{z}^{j\ell}_t \in \mathbf{R}^H$, the previous time step's hidden state $\boldsymbol{h}^{j\ell}_{t-1} \in \mathbf{R}^H$, and

the cell state $\boldsymbol{c}_{t-1}^{j\ell} \in \mathbf{R}^H$. The LSTM cell produces a hidden output $\boldsymbol{h}_t^{j\ell}$, which is then passed through three different feedforward layers:

$$\boldsymbol{p}_t^{j\ell} = \text{FeedForward}^{P\ell}(\boldsymbol{h}_t^{j\ell}) \tag{45}$$

$$\boldsymbol{q}_t^{j\ell} = \text{FeedForward}^{Q\ell}(\boldsymbol{h}_t^{j\ell}) \tag{46}$$

$$\boldsymbol{u}_t^{j\ell} = \text{FeedForward}^{U\ell}(\boldsymbol{h}_t^{j\ell}) \tag{47}$$

Each of the feedforward layers consists of two linear projections with a GELU activation after the first linear projection:

$$\text{FeedForward}(\boldsymbol{h}) = \boldsymbol{W}^{FF_2}\,\text{GELU}(\boldsymbol{W}^{FF_1}\,\boldsymbol{h}) \tag{48}$$

The GELU activation function is a stochastic variant of ReLU that has been shown to outperform ReLU in sequence-to-sequence models [**?**]. It is defined as $\text{GELU}(x) = x\,\Phi(x)$, where Φ is the standard Gaussian cumulative distribution function.

The first output $\boldsymbol{p}_t^{j\ell}$ is a component of the projected time series captured by Block_ℓ. Subsequent blocks depend on the residual value of the projected time series after removing this component:

$$\boldsymbol{z}_t^{j\ell+1} = \boldsymbol{z}_t^{j\ell} - \boldsymbol{p}_t^{j\ell} \tag{49}$$

The second output $\boldsymbol{q}_t^{j\ell}$ is Block_ℓ's contribution to the forecast of the next time step. The final forecast representation of the recurrent component will be the sum of all the blocks

$$\hat{\boldsymbol{q}}_{t+1}^{jR} = \sum_{\ell=1}^{L} \boldsymbol{q}_t^{j\ell} \tag{50}$$

where $\hat{\boldsymbol{q}}_{t+1}^{jR} \in \mathbf{R}^H$. We then project this into $\mathbf{R}^D$ to get the forecast contribution from the recurrent component, i.e. the first term in Equation (42):

$$\hat{\boldsymbol{v}}_{t+1}^{jR} = \boldsymbol{W}^R\,\hat{\boldsymbol{q}}_{t+1}^{jR} \tag{51}$$

2.2.2 Flow aggregation component

The flow aggregation component models the influence between the time series of neighboring nodes in the network. This component takes as input time-dependent embeddings from the recurrent component of each node in the neighborhood, and produces as output the second term in Equation (42). Each embedding summarizes the time series of the corresponding node up to the current time. Let $\boldsymbol{u}_t^j$ be the embedding of the ego node $\boldsymbol{v}^j$ at time step t, formed by summing the node vectors $\boldsymbol{u}_t^{j\ell}$ over all L blocks:

$$\boldsymbol{u}_t^j = \sum_{\ell=1}^{L} \boldsymbol{u}_t^{j\ell} \tag{52}$$

In the IMPUTATION setting, the set of embeddings of all neighbors of the ego at time $t+1$ is

$$\mathcal{U}_{t+1}^{\boldsymbol{v}^j} = \{\boldsymbol{u}_{t+1}^i \mid i \text{ s.t. } \boldsymbol{v}^i \in \mathcal{N}_{t+1}(\boldsymbol{v}^j)\} \tag{53}$$

while in the FORECAST setting, we simply replace the ground truth $\boldsymbol{u}_{t+1}^i$ with the forecast $\hat{\boldsymbol{u}}_{t+1}^i$. We now project the ego's embedding into the query space,

$$\boldsymbol{u}_t^{Qj} = \boldsymbol{W}^Q \, \boldsymbol{u}_t^j \tag{54}$$

and the neighbors' embeddings into the key and value space,

$$\boldsymbol{u}_{t+1}^{Ki} = \boldsymbol{W}^K \, \boldsymbol{u}_{t+1}^i \qquad \forall i \text{ s.t. } \boldsymbol{v}^i \in \mathcal{N}_{t+1}(\boldsymbol{v}^j) \tag{55}$$

$$\boldsymbol{u}_{t+1}^{Vi} = \boldsymbol{W}^V \, \boldsymbol{u}_{t+1}^i \qquad \forall i \text{ s.t. } \boldsymbol{v}^i \in \mathcal{N}_{t+1}(\boldsymbol{v}^j) \tag{56}$$

The aggregated embedding $\tilde{\boldsymbol{u}}_{t+1}^j$ is then the weighted sum of the values with a GELU activation,

$$\tilde{\boldsymbol{u}}_{t+1}^j = \text{GELU}\Big(\sum_i \lambda_i \, \boldsymbol{u}_{t+1}^{Vi} \Big) \tag{57}$$

where the weights λ_i, called *attention scores*, are computed from the dot product between the query and the keys, followed by a softmax. Note that the ego node is not included in the aggregation; instead it is added separately,

$$\hat{\boldsymbol{u}}_{t+1}^j = \boldsymbol{W}^E \boldsymbol{u}_t^j + \boldsymbol{W}^N \tilde{\boldsymbol{u}}_{t+1}^j \tag{58}$$

which is then projected down to $\mathbf{R}^D$,

$$\hat{\boldsymbol{v}}_{t+1}^{jA} = \boldsymbol{W}^A \, \hat{\boldsymbol{u}}_{t+1}^j \tag{59}$$

The vector $\hat{\boldsymbol{v}}_{t+1}^{jA}$ is the forecast contribution from the flow aggregation component, i.e. the second term in Equation (42).

We call the full model with multi-head attention *Radflow*. Note that the flow aggregation component and the recurrent component are decoupled. Thus we can easily substitute the multi-head attention with another node aggregation method. In particular, if we replace Equation (57) with a simple arithmetic average of the neighbors,

$$\tilde{\boldsymbol{u}}_{t+1}^j = \frac{1}{|\mathcal{N}_{t+1}(\boldsymbol{v}^j)|} \sum_i \boldsymbol{u}_{t+1}^i \tag{60}$$

we would obtain the original formulation of GraphSage [48]. We call the model that uses Equation (60) instead of Equation (57) *Radflow-GraphSage*. In addition to adopting Equation (60), a further simplification is to remove the linear projection in Equation (58) when adding the ego's embedding with its neighbors'. Let us call this variant *Radflow-MeanPooling*.

2.2.3 Relationship with existing models

GAT Our multi-head attention neighborhood aggregation is similar to the Graph Attention Network (GAT) [89]. To compute the attention score in GAT, we first need to concatenate the ego node's embedding with the neighbor's, and then feed the result through a single feedforward network followed by a LeakyReLU. In contrast, we revert back to the original multi-head attention [88] where we compute the attention scores with a simple dot product. We also add zero attention, in which a node has the option not to attend to any neighbor.

N-BEATS The process of feeding residuals of time series into deep network layers is inspired by N-BEATS. However, N-BEATS takes residuals from the raw scalar observations, whereas our approach calculates the residuals from the vector-valued projections of the time series, as shown in Equation (43). Moreover, N-BEATS is not easily adapted to the dynamic graph setting since it does not produce embeddings that depend on time. N-BEATS treats the forecasting task as a multivariate regression problem, where every step can see every other step in the history. This allows us to obtain an embedding for the whole series but not for an individual step. In our proposed architecture, the node vector $\boldsymbol{u}_t^{j\ell}$ is used to construct the time-dependent embedding of each step, as shown in Equation (52).

Transformers In the last few years, transformers [88] have become the sequence model of choice in the NLP domain. Despite their success in NLP tasks, little progress has been made with time series forecasting. Most recently, [93] designed a transformer to forecast flu cases, but their model provides only marginal improvement over the LSTM baseline. Our preliminary investigation indicated that LSTMs perform better than transformers in the time series setting. We hypothesize that the strict temporal ordering of the LSTM can encode time series more naturally; while text, which often has a latent tree structure, is more naturally encoded by the transformer with its attention mechanism and position encodings.

Non-neural aggregation The most relevant non-neural aggregation approach is ARNet [96], a forecasting model for scalar-valued time series in which the prediction is computed as:

$$\hat{v}_t^j = \sum_{k=1}^{p} \alpha_k^j v_{t-k}^j + \sum_{v^i \in \mathbf{N}(\boldsymbol{v}^j)} \beta^{ij} v_t^i \tag{61}$$

where the first term is an autoregressive model of order $p = 7$ (days) and the second term models the network effect. The learnable parameters β^{ij} can be interpreted as the edge weight that controls the proportion of views propagating from node i to node j. Although ARNet is simple with a straightforward interpretation, the model assumes that the network is static. Furthermore [96] only evaluated on the IMPUTATION setting and not on the FORECAST setting where the future observations are unknown.

Neural aggregation The closest model to ours is T-GCN [101], where a modified GRU cell does a graph convolution before computing the update and reset gates. Unlike Radflow which has been implemented to fetch subgraphs from disk and only compute the network information once after the final LSTM layer, T-GCN requires the entire network to be in memory and aggregates the network at every time step in every layer. Therefore, T-GCN does not scale to larger datasets due to both space and time complexity. Our proposed architecture, on the other hand, can easily handle dynamic networks of hundreds of thousands of nodes.

2.3 Predicting time series networks

The empirical validation of Radflow is carried out on two small static networks—LOS-LOOP and SZ-TAXI [101]; and two large-scale dynamic networks—VEVOMUSIC [96] and WIKITRAFFIC.

Out of these, WIKITRAFFIC is a new dataset that we have collected and it is the largest dynamic network of time series to date.

2.3.1 Evaluation settings

Evaluation is done by predicting view counts of the final 7 days on VEVOMUSIC and the final 28 days on WIKITRAFFIC. For LOS-LOOP and SZ-TAXI, we predict the speeds in the final hour (the final 12 steps in LOS-LOOP and the final 4 steps in SZ-TAXI). As stated in Section 2.1, we consider both the FORECAST and IMPUTATION settings for the large datasets. On VEVOMUSIC, we evaluate on two different networks: the full dynamic network which we call VEVOMUSIC (dynamic), and the static network which we call VEVOMUSIC (static). To construct the static version, [96] used a majority smoothing method to remove edges that occur only briefly and made the remaining edges persistent in all time steps. Their best model ARNet was only evaluated on this static network in the IMPUTATION setting. On WIKITRAFFIC, we consider two networks: one is a network of univariate time series of view counts, while the other is a network of bivariate time series where desktop and non-desktop traffic are split.

Following prior forecasting work [65, 66], our main evaluation metric will be the Symmetric Mean Absolute Percentage Error for forecast horizon F:

$$\text{SMAPE-}F = \frac{100}{\mathcal{T}FD} \sum_{j=1}^{\mathcal{T}} \sum_{t=1}^{F} \sum_{d=1}^{D} \frac{\left|v_{td}^{j} - \hat{v}_{td}^{j}\right|}{\frac{1}{2}\left(\left|v_{td}^{j}\right| + \left|\hat{v}_{td}^{j}\right|\right)} \tag{62}$$

where $\mathcal{T}$ is the number of samples in the test set, F is the forecast horizon, D is the dimension of the time series, and $\hat{v}_{td}^{j}$ is the forecast value of the ground truth v_{td}^{j}. SMAPE is interpretable with an upper bound of 200 and a lower bound of 0. It is scale-independent, ensuring that prediction errors will be considered relative to the magnitude of the sequence. This is important because it prevents nodes with a large number of views from dominating the evaluation measure. A lower SMAPE corresponds to a better fit, with it being 0 if and only if the prediction matches the ground truth perfectly. For the two small networks of univariate time series, LOS-LOOP and SZ-TAXI, we additionally report the Root Mean Square Error,

$$\text{RMSE-}F = \sqrt{\frac{1}{\mathcal{T}F} \sum_{j=1}^{\mathcal{T}} \sum_{t=1}^{F} \left(v_{t}^{j} - \hat{v}_{t}^{j}\right)^{2}} \tag{63}$$

and the Mean Absolute Error,

$$\text{MAE-}F = \frac{1}{\mathcal{T}F} \sum_{j=1}^{\mathcal{T}} \sum_{t=1}^{F} \left|v_{t}^{j} - \hat{v}_{t}^{j}\right| \tag{64}$$

similar to what was done in the T-GCN paper [101].

2.3.2 Model variants

We compare 8 time series baselines, 7 variants of networked time series, and 7 more variants of Radflow in an ablation study. Results of the following 8 time series baselines are in Table 2 and Table 3.

Table 1: Hyperparameters of **Radflow-NoNetwork** (8) and Radflow (15) on WIKITRAFFIC (univariate). We calibrate the hidden size to ensure that the number of parameters of all models are within 5% of each other. See the Appendix [86] for the hyperparameters of other model variants.

	0 Hops	1 Hop	2 Hops
Number of parameters	1,589,762	1,608,464	1,576,52
LSTM hidden size	128	116	112
Feedforward hidden size	128	116	112
Number of LSTM layers (L)	8	8	8
Dropout probability	0.1	0.1	0.1
Number of attention heads	-	4	4
Backcast (seed) length	112	112	112
Forecast length	28	28	28

(1) **Copying Previous Step:** We use the final observation before the test period as the prediction. This is the final day in VEVOMUSIC and WIKITRAFFIC, the final 5 minutes in LOS-LOOP, and the final 15 minutes in SZ-TAXI.

(2) **Copying Previous Week:** Since we observe weekly seasonality in both VEVOMUSIC and WIKITRAFFIC, a stronger baseline is to copy observations in the final week just before the test period and use them as the predictions.

(3) **AR:** The autoregressive (AR) model used by [96] for the static VEVOMUSIC network.

(4) **Seasonal ARIMA:** We train an ARIMA$(p, d, q)(P, D, Q)_m$ model separately for each time series, where p, d, q, P, D, Q are the AR, difference, MA, seasonal AR, seasonal difference, and seasonal MA terms, respectively. These are learned automatically using the `pmdarima` package. The number of periods in a season m is set to 7 days for VEVOMUSIC and WIKITRAFFIC.

(5) **Individual LSTMs:** The LSTM baseline used by [96]. It is trained separately for each time series, with no weight sharing across network nodes.

(6) **LSTM:** The standard LSTM model with weight sharing. Unlike variant (4), this method uses only one set of LSTM weights for the entire dataset.

(7) **N-BEATS:** The neural regression with residual stacking from [74]. The implementation consists of eight stacks, each containing one generic block. A generic block internally uses four fully-connected layers, followed by a fork into the forecast and backcast space. For bivariate WIKITRAFFIC, we train two separate N-BEATS models.

(8) **Radflow-NoNetwork:** Radflow with only the recurrent component, i.e. first term $\hat{\boldsymbol{v}}_t^{jR}$ in Equation (42). It does not take any contribution from the network.

Results for the forecasting models which use the network structure are presented in Table 3 and 4.

(9) **T-GCN:** The model proposed by [101] that uses a modified GRU cell to aggregate nodes. Since T-GCN is not scalable to larger datasets (see Section 2.2.3), we only compare Radflow against T-GCN on LOS-LOOP and SZ-TAXI [101].

(10) **ARNet:** The state-of-the-art model [96] for the VEVOMUSIC (static) dataset in the IMPUTATION setting (see Section 2.2.3).

(11) **LSTM-MeanPooling:** The same architecture as (6) but with mean-pooling node aggregation, using the hidden output of the final LSTM layer as a node's representation.

(12–15) **Radflow:** Variants of our proposed architecture with different network aggregation techniques: a simple mean (12), GraphSage (13), Graph Attention Network (14), and our full Radflow model with multi-head attention (15). Table 1 outlines the hyperparameters used in the full model.

Finally we conduct an ablation study to test the key components of our architecture (Table 5). Starting with the best model (15), we substitute one component with an alternative:

(16–20) **Radflow with other node embeddings:** As shown in Section 2.2, the LSTM cell contains a hidden state $\boldsymbol{h}$ which is used to produce three vectors, $\boldsymbol{p}$, $\boldsymbol{q}$, and $\boldsymbol{u}$. Instead of having a separate output $\boldsymbol{u}$ to represent a node, we could alternatively reuse the cell's hidden state $\boldsymbol{h}$ (16), the backcast representation $\boldsymbol{p}$ (17), or the forecast representation $\boldsymbol{q}$ (18). We could also concatenate different representations, such as $[\boldsymbol{h}; \boldsymbol{p}]$ (19) or $[\boldsymbol{h}; \boldsymbol{p}; \boldsymbol{q}]$ (20).

(21) **Radflow with no final projection:** We ignore the linear projection in Equation (58) and add the ego node's embedding to its neighbors' directly.

(22) **Radflow with one attention head:** This final variant tests the effect of having only one attention head instead of the default four in the full model.

2.3.3 Data preprocessing and training details

Web-scale time series observations often vary greatly in scale. An unpopular page might get zero views, while a popular page might receive millions of visits daily. To ensure similar scaling, both the inputs and outputs of our models are log-transformed time series. Outputs are exponentiated before computing SMAPE, RMSE, and MAE. Missing views are imputed by propagating the last valid observation forward. We do not apply any other preprocessing techniques to the time series, such as trend or seasonality removal.

Table 1 shows key hyperparameters of Radflow. We trained all models on the SMAPE objective using the Adam optimizer [57] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$. We set the weight decay factor to 10^{-4} and decouple it from the learning rate [64]. We warm up the learning rate to 10^{-4} in the first 5,000 steps and then linearly decay it afterward for 10 epochs, each of which consists of 10,000 steps. We clip the gradient norm at 0.1. All our models are implemented in Pytorch [75]. For a fair comparison, we fix the number of layers of all variants to eight and ensure that size of all variants are within 5% of each other.

Table 2: Performance of time series models with no network information. We report mean SMAPE-7 on VEVOMUSIC (22 Oct 18 – 2 Nov 18) and mean SMAPE-28 on WIKITRAFFIC (3 Jun 20 – 30 Jun 20). Rows are numbered according to Sec 2.3.2. See the Appendix [86] for statistical significance tests.

	VEVO MUSIC	WIKITRAFFIC (univariate)	WIKITRAFFIC (bivariate)
(1) Copying Previous Step	14.0	22.5	26.8
(2) Copying Previous Week	10.3	21.0	25.4
(3) AR [96]	10.2	-	-
(4) Seasonal ARIMA	9.67	19.6	22.8
(5) Individual LSTMs [96]	9.99	-	-
(6) LSTM	8.68	16.6	20.4
(7) N-BEATS	8.64	16.6	20.3
(8) Radflow-NoNetwork	**8.42**	**16.1**	**19.4**

Table 3: FORECAST performance on the static traffic networks. On LOS-LOOP, we report mean SMAPE-12, RMSE-12, and MAE-12. On SZ-TAXI, we report mean SMAPE-4, RMSE-4, and MAE-4.

	LOS-LOOP			SZ-TAXI		
	SMAPE	RMSE	MAE	SMAPE	RMSE	MAE
(1) Copying Previous Step	3.92	3.40	2.39	**45.8**	4.32	2.81
(8) Radflow-NoNetwork	3.60	3.23	2.18	80.2	3.99	3.06
(9) T-GCN [101]	3.97	3.42	2.41	80.5	6.27	3.52
(15) Radflow	**3.50**	**3.11**	**2.11**	77.5	**3.36**	**2.51**

2.3.4 Computational costs

Training time VEVOMUSIC experiments were trained on a Titan V GPU and WIKITRAFFIC experiments were trained on a Titan RTX GPU. The Titan RTX has twice the memory of the Titan V and is needed to train the two-hop Radflow on WIKITRAFFIC. All pure time series models converge very quickly, taking no more than three hours to train. Models with one-hop aggregation take up to 17 hours to train, while models with two-hop aggregation can take up to two days. We pick the model from the epoch with the lowest SMAPE score on the validation set as our best model.

Efficient computation of graphs Unlike previous approaches such as T-GCN, our models do not require the whole graph to be in memory during training. Instead we store the graph in the HDF5 format and only load one batch at a time directly from disk.

Neighborhood sampling To keep the computation tractable, we devise an importance-based neighborhood sampling technique. Instead of the common uniform sampling that was, for example, used by [48], we propose a two-stage approach to select neighbors. First we assign each

Table 4: Performance of models with network information. We report mean SMAPE-7 on VEVOMUSIC (22 Oct 18 – 2 Nov 18) and mean SMAPE-28 on WIKITRAFFIC (3 Jun 20 – 30 Jun 20). Rows are numbered according to Sec 2.3.2. Bold numbers indicate the best model(s) within a column. Refer to the Appendix [86] for the p-values from the dependent t-test for paired samples between models with similar performance.

	VEVOMUSIC (static)				VEVOMUSIC (dynamic)				WIKITRAFFIC (univariate)				WIKITRAFFIC (bivariate)			
	Forecast		Imputation		Forecast		Imputation		Forecast		Imputation		Forecast		Imputation	
	1H	2H	1H	2H	1H	2H	1H	2H	1H	2H	1H	2H	1H	2H	1H	2H
(10) ARNet [96]	-	-	9.02	-	-	-	-	-	-	-	-	-	-	-	-	-
(11) LSTM-MeanPooling	8.60	8.67	8.14	8.13	8.80	9.03	7.91	7.90	16.8	16.7	15.5	15.2	20.2	19.9	19.2	18.9
(12) Radflow-MeanPooling	8.34	8.44	7.82	7.81	8.42	**8.32**	7.74	7.61	16.5	16.5	15.1	15.1	19.8	20.2	18.5	18.6
(13) Radflow-GraphSage	8.39	**8.37**	7.78	7.64	8.43	8.46	7.46	**7.27**	**15.9**	16.7	14.7	15.0	**19.3**	19.9	**18.2**	**18.4**
(14) Radflow-GAT	8.52	8.50	7.88	7.74	8.43	8.39	7.44	7.28	16.2	**16.0**	15.0	15.2	19.5	19.7	18.3	18.6
(15) Radflow	**8.33**	8.39	**7.67**	**7.63**	**8.37**	8.45	**7.32**	**7.27**	16.2	**16.0**	**14.5**	**14.8**	19.9	**19.6**	18.3	18.5

Table 5: Ablation study on the key components of Radflow on one-hop VEVOMUSIC networks.

	VEVOMUSIC (static)	VEVOMUSIC (dynamic)
(15) Radflow	**7.67**	**7.32**
(16) Radflow ($\boldsymbol{h}$ as embeddings)	7.78	7.51
(17) Radflow ($\boldsymbol{p}$ as embeddings)	7.77	7.49
(18) Radflow ($\boldsymbol{q}$ as embeddings)	7.84	7.35
(19) Radflow ($[\boldsymbol{h}; \boldsymbol{p}]$ as embeddings)	7.75	7.39
(20) Radflow ($[\boldsymbol{h}; \boldsymbol{p}; \boldsymbol{q}]$ as embeddings)	7.76	7.38
(21) Radflow (no final projection)	7.80	7.33
(22) Radflow (one attention head)	7.77	7.43

neighbor a score $\frac{\sum_d v_{td}^j}{\text{outdegree}(\boldsymbol{v}_t^j)+1}$. This score is the total view count of the neighbor at time step t, normalized by the outdegree of the neighbor at that time. A self-loop is added to avoid division by zero. Intuitively, a neighbor with a larger number of views will have a greater influence, but the influence will be more diffuse if that neighbor has many outlinks. Using these scores, we remove neighbors in the bottom 10th percentile in the neighborhood of each ego node, which reduces noise induced by aggregation.

In the second stage, we sample four neighbors during training with probability proportional to the number of time steps that the neighbor appears in the backcast period. During evaluation, we find using all nodes to be computationally infeasible due to the large data size. Thus for each ego node, we choose only the 16 most frequently appearing neighbors in the one-hop setting and the top eight in the two-hop setting.

2.4 Selected results

Radflow is evaluated on four datasets. Two are on urban traffic consisting of several hundred nodes; two are large-scale datasets—VEVOMUSIC containing 61K videos [96] and a newly curated WIKITRAFFIC dataset containing 366K pages and 22M dynamic links. On both VEVOMUSIC and WIKITRAFFIC, Radflow without network information is consistently better than the comparable N-BEATS [74]. Among models with network information, Radflow variants perform the best in both

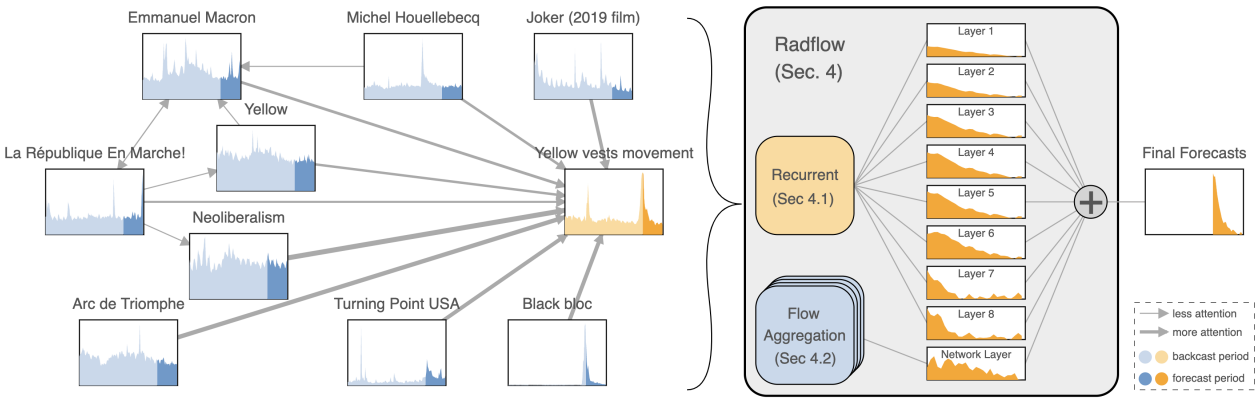


Figure 3: Overview of the Radflow model, centered around the WIKITRAFFIC subgraph of *Yellow vests movement* (a social movement in France since 2018, shown on the left). Each node is a page with a time series of view counts, shown in its individual mini-panel. Edge strengths correspond to average attention scores in the forecast period. The final forecast is produced by summing over the output from eight layers of the recurrent component and the network flow aggregation component (shown on the right). Radflow correctly predicts a sharp drop in traffic volume for *Yellow vests movement* from 3 June 2020 to 30 June 2020, due to information from neighboring nodes such as *Black bloc* and *Turning Point USA*.

imputation and forecasting tasks. In particular, Radflow outperforms state-of-the-art ARNet [96] by 19% in SMAPE score on VEVOMUSIC. We find that the layers in the recurrent component capture different seasonality and trends, while attention over the network captures the time-varying influence from neighboring nodes. Figure 3 illustrates the task of predicting 28 days of view counts on *Yellow vests movement*, based on the historical traffic of that page and the traffic of the neighboring pages. Radflow correctly predicts the sharp drop that is observed during the test period.

We first discuss prediction performances of different model variants (Sections 2.4.1 and 2.4.2, Tables 2 to 5). When applicable, we report in parentheses the p-value (denoted as p) from the dependent t-test for paired samples. All differences discussed in this section are statistically significant. For more detailed significance tests, see the Appendix [86]. We then present a visual interpretation of different layers in the recurrent component (Section 2.4.3), followed by insights provided by the network aggregation component (Section 2.4.4). Finally, we present two preliminary studies on the potential applications of models like Radflow: the robustness of predictions when the network is not fully observed (Section 2.4.5), and the relationship between traffic surges on nodes and their attention scores (Section 2.4.5).

2.4.1 Forecasting and imputation performance

FORECASTING **without networks** Table 2 summarizes the comparison between Radflow-NoNetwork and the corresponding time series forecasting baselines (1–7). The LSTM variant (6) outperforms both AR (3) and Seasonal ARIMA (4), showing the robust performance of flexible neural models. Furthermore, it also outperforms models trained on individual time series (3, 4, 5), highlighting the advantage of using large amounts of training data. N-BEATS (7) outperforms LSTM (6) by a small margin of 0.04 SMAPE ($p = 6\,\mathrm{e}{-3}$) on VEVOMUSIC, while Radflow-NoNetwork outperforms all other baselines, showing promises in combining the recurrent structure with the residual stacking

idea in our architecture.

Additionally, Radflow-NoNetwork outperforms ARNet (10), the state-of-the-art for VEVOMUSIC that uses network information (Table 4), indicating that having the right model outweighs having more information for this task.

FORECASTING **with networks** Tables 3 and 4 summarize performances of models (9–15) on the four networked time series datasets. Our full model (15) outperforms T-GCN in LOS-LOOP by a non-trivial margin on all metrics. SZ-TAXI is more noisy and no model is able to beat the SMAPE from copying the previous step. This is because SZ-TAXI contains many consecutive zero measurements, which the copying baseline is able to take advantage of. On non-zero test measurements, Radflow is able to outperform all other variants. See the Appendix [86] for these results.

Across all eight FORECASTING settings on VEVOMUSIC and WIKITRAFFIC, the top-performing models are all Radflow variants. Compared to Radflow-NoNetwork (8), incorporating one-hop neighbors improves the SMAPE score on VEVOMUSIC from 8.42 to 8.33 ($p = 2\,\mathrm{e}{-33}$). On WIKITRAFFIC (univariate), using one-hop GraphSage (13) improves SMAPE from 16.1 to 15.9 ($p = 2\,\mathrm{e}{-21}$). This confirms the recently reported network effects in the YouTube and Wikipedia traffic [96, 104]. Note that WIKITRAFFIC is much larger and more diverse, making consistent improvements harder, hence the smaller effect size compared to VEVOMUSIC.

Finally, compared to the smoothed static network, using the dynamic network in VEVOMUSIC only improves the performance for some model variants. This is because the static graph was constructed from the smoothing of edges (i.e. uncommon edges were removed), which reduces noise in the absence of ground-truth views from the neighbors. More generally, although Radflow is designed to handle dynamic graphs, it is impossible to know a priori whether dynamic edges will improve the prediction performance in a given data domain.

IMPUTATION Having ground-truth observations of neighboring nodes during the forecast period leads to substantially better SMAPE scores for all models. Models that perform well in the FORECAST setting also perform well in the IMPUTATION setting, with Radflow (15) achieving the highest SMAPE in six of the eight IMPUTATION settings. The performance gain going from one hop to two hops is more evident in the IMPUTATION setting. For example, there is a boost from 7.67 to 7.63 ($p = 1\,\mathrm{e}{-8}$) on the static VEVOMUSIC, and from 7.32 to 7.27 ($p = 3\,\mathrm{e}{-14}$) on the dynamic VEVOMUSIC. Compared to the previous state-of-the-art ARNet [96], our best model achieves a SMAPE score that is 19% better (from 9.02 to 7.27).

On WIKITRAFFIC, using two-hop neighbors uniformly lowers the performance, while the training time more than doubles. Compared to VEVOMUSIC, each Wikipedia page has many more links, most of which would be ignored by the reader. The network effect in WIKITRAFFIC is thus weaker than VEVOMUSIC (see Figure 5), leading to more noise being introduced in the two-hop setting.

2.4.2 Radflow components

Among the variants with attention, the weighted multi-head attention of Radflow yields the best performance in 10 out of 16 tasks across FORECAST and IMPUTATION, while GraphSage, which puts an equal weight on all neighbors, is the second best (best in 6 out of 16 tasks). This indicates that simpler attention mechanisms (inner-product attention in Radflow and node averaging in

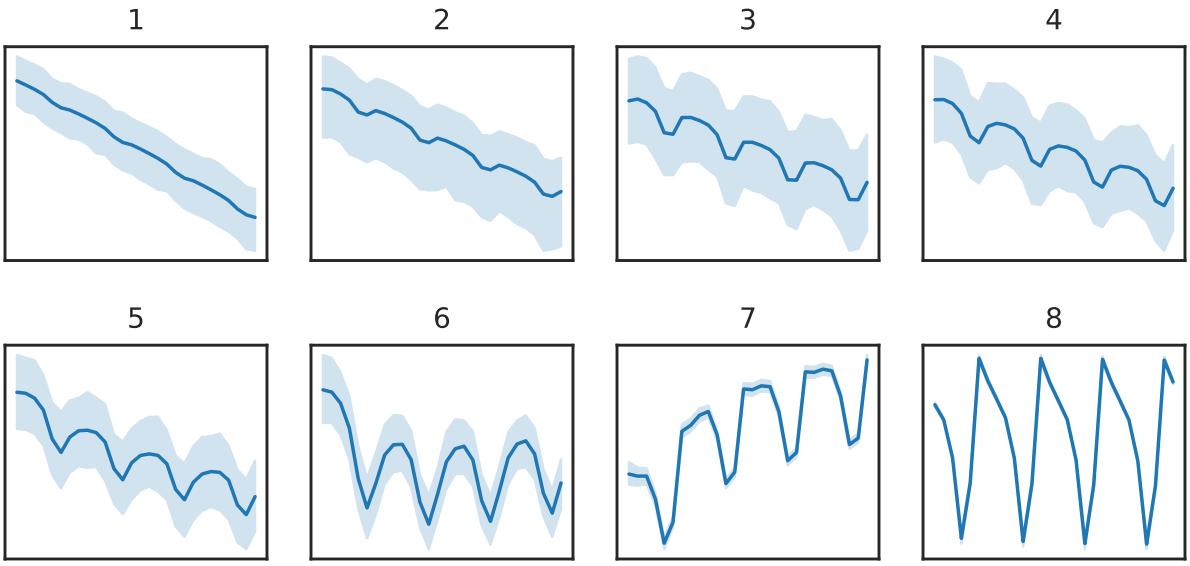


Figure 4: Average prediction over all test pages from each layer of Radflow-NoNetwork on the 28-day test period of WIKITRAFFIC. The light shade corresponds to the 95% confidence interval. Different layers capture different seasonal and trend patterns. See Section 2.4.3.

Radflow-GraphSage) are preferred over more complex ones (Radflow-GAT), but it should not be too simple (Radflow-MeanPooling).

Table 5 presents an ablation study that tests other key components in Radflow. Overall, we find that having more than one attention head helps (15 vs 22), so does a linear projection before node aggregation (15 vs 21). It is also preferable to have a separate projection on the output of the LSTM cell to obtain the node embeddings $\boldsymbol{u}$, than to re-use the hidden state $\boldsymbol{h}$, the backcast representation $\boldsymbol{p}$, or the forecast representation $\boldsymbol{q}$ (15 vs 16–20). All differences are statistically significant.

2.4.3 Layered decomposition of time series

The recurrent component of Radflow is decomposable into $L = 8$ layers, via Equations (50) and (51). Figure 4 shows the layer-wise contribution to the forecast from Radflow-NoNetwork. The results are averaged over all 2,434 test pages in WIKITRAFFIC over the 28-day test period, and re-scaled to the same range for readability. We observe that component 8 encodes strong weekly seasonality that is consistent across all test pages (with a small confidence interval). Components 1–5 encode varying levels of a decreasing trend, whereas component 7 encodes an increasing trend. Overall, weekly seasonality is visible in all components except the first, confirming the common intuition that representations learned via neural networks are often over-complete.

2.4.4 Quantifying effects of the network

Network contribution Radflow can identify settings where the network information becomes important for the final predictions. Figure 5 shows the contribution to the final forecast made by the network component (the second term of Equation (42)). In both VEVOMUSIC and WIKITRAFFIC,

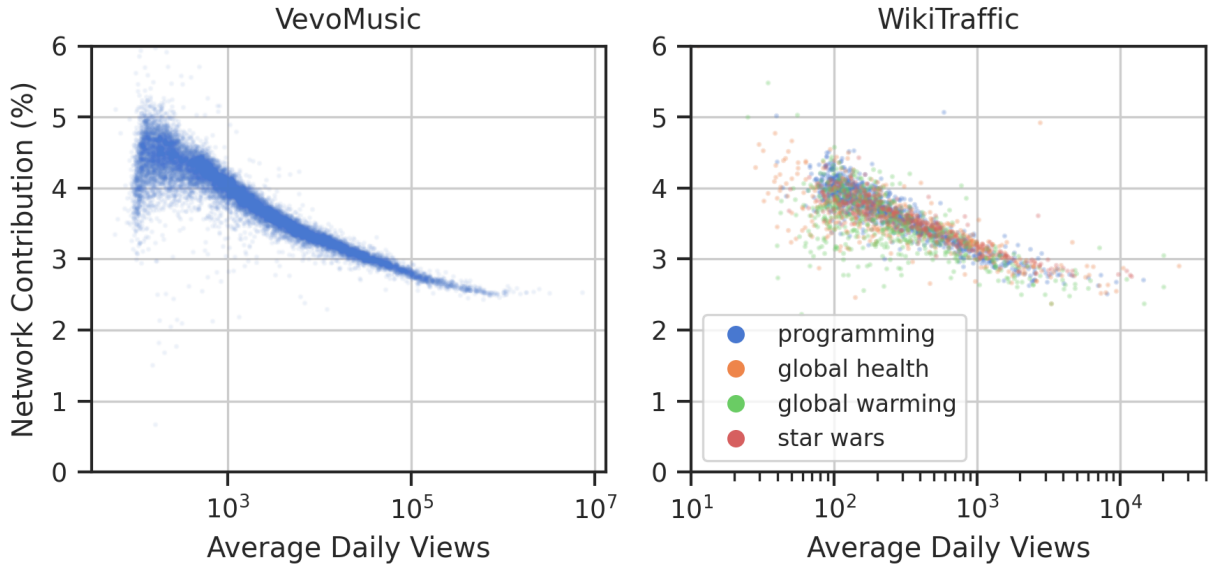


Figure 5: The contribution from the network on the final forecasts. The network component of less popular nodes is larger. For nodes with similar daily view counts, VEVOMUSIC (left) exhibits a stronger network effect than WIKITRAFFIC (right). See Section 2.4.4.

there is an approximately inverse linear relationship between the network contribution and the log of average daily views, indicating that forecasts of less popular nodes rely more heavily on the network. For nodes with a similar number of daily views (take, for example, 10^3 views per day), VEVOMUSIC exhibits a stronger network effect ($\sim$4%) than WIKITRAFFIC ($\sim$3%), indicating that user behaviors on YouTube recommender links and Wikipedia hyperlinks are different, warranting further investigation (e.g., with user-level data).

Visualising attention flow The attention scores in Radflow capture some information about the flow of traffic. In particular, we observe that the model pays more attention to a neighboring node on days that have a spike in traffic (see the Appendix [86] for a specific example). This motivates us to visualize the attention flow among WIKITRAFFIC nodes in an interactive web app [82]. In this visualization, nodes are represented in a graph (bottom panel) and as time series (top panel). Edge weights are attention scores from Radflow multiplied by the traffic on the source node.

Figure 6 is a screenshot for the subgraph centered around the Wikipedia page of *Kylo Ren*, a character in the Star Wars series played by the actor *Adam Driver*. From the figure, we observe that the time series of the two nodes (*Kylo Ren* in blue, *Adam Driver* in pink) have synchronized spikes at the same time as the release of major Star Wars movies. Furthermore, there appears to be more traffic flowing from *Kylo Ren* to *Adam Driver* than the other way, indicated by the thickness of the edges.

2.4.5 What if . . .

Data is missing? We consider the robustness of Radflow to missing data. This setting is relevant when collecting and observing data from all nodes is very costly (e.g., sites spread out over large

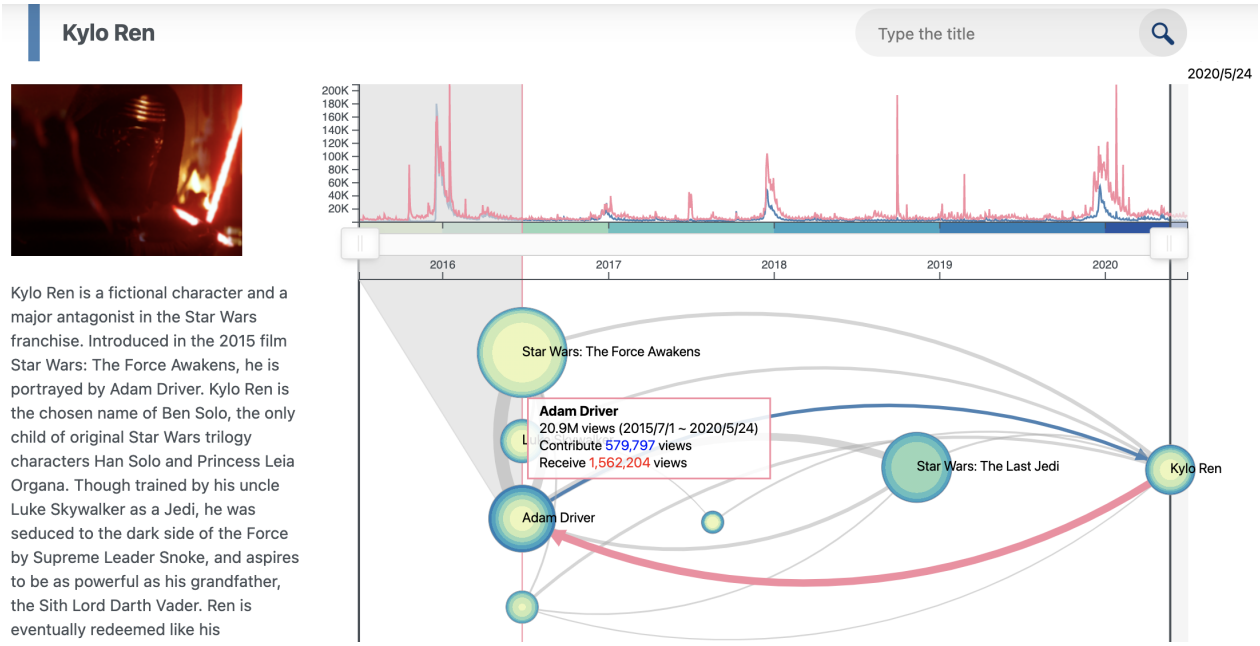


Figure 6: An interactive web app [82] for attention flow, available at https://attentionflow.ml. Shown here is the subgraph for the Wikipedia page *Kylo Ren*, a character in *Star Wars* played by *Adam Driver*. The thickness of an edge is determined by the product between the attention score and the traffic volume of the source node. See Section 2.4.4.

geographical areas) or when nodes are simply unavailable (e.g., sites whose data are proprietary). To this end, Figure 7 shows our evaluation of Radflow on the IMPUTATION task for VEVOMUSIC with a percentage of time series values (left pane) or edges (right pane) deleted at random. As more nodes become missing, the performance of Radflow decays at a much slower rate than Radflow-NoNetwork. This indicates that Radflow is effective in imputing and mitigating missing values from other neighbors. Similarly, Radflow is relatively robust to missing edges. With 40% of edges missing, the performance of the two-hop model only drops by 1% in SMAPE. Even with 80% of edges missing, Radflow is still substantially better than Radflow-NoNetwork.

Node traffic doubles? In addition to the example of an actual traffic spike in Figure 3, we further perform an evaluation to visualize the effects of hypothetical sudden changes in node traffic. This scenario is useful for planning resources such as edge network caching for mobile videos, or for estimating economic demands associated with nodes such as advertising.

For each test node in the IMPUTATION task on WIKITRAFFIC, we pick one neighbor and double its views on one forecasting day. Figure 8 (left) shows that as a page becomes more popular, the attention on it uniformly decreases, indicating that attention tends to dampen node traffic spikes instead of amplifying them. More evidently, we can see from Figure 8 (right) that a higher attention score *is positively correlated with a large effect on the ego node's traffic*. This indicates that despite the non-linear relationship between attention scores and network component in Section 2.2, one could qualitatively infer the traffic flow from one page to another using attention scores.

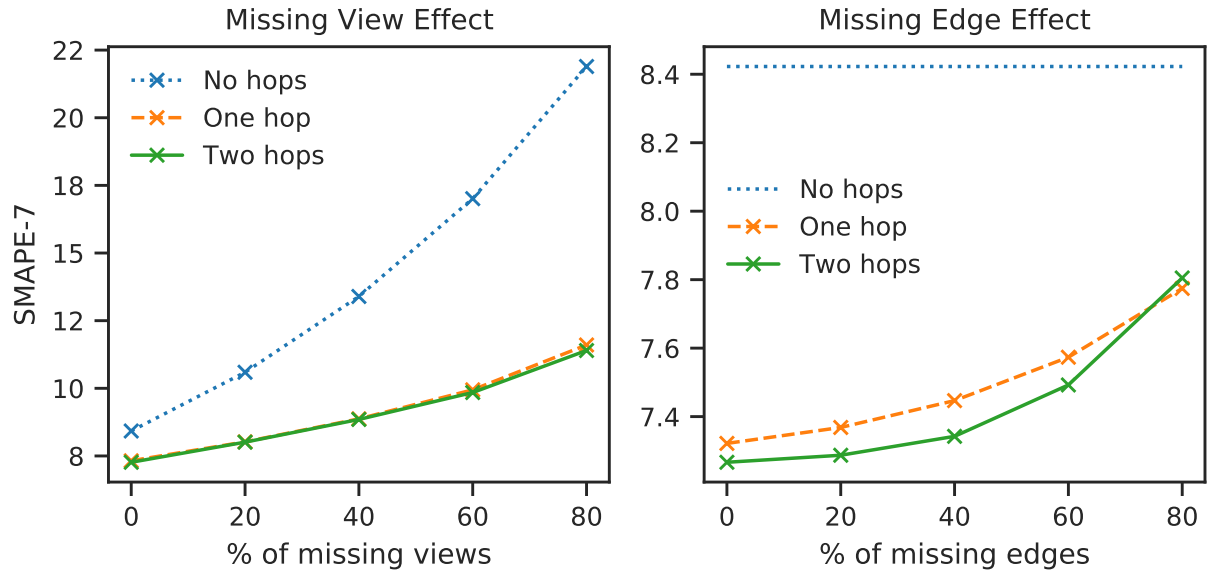


Figure 7: The effect of missing view counts (left) and missing edges (right) on VEVOMUSIC using Radflow. As we delete more data, Radflow's performance degrades at a much slower rate than Radflow-NoNetwork (in blue). See Section 2.4.5.

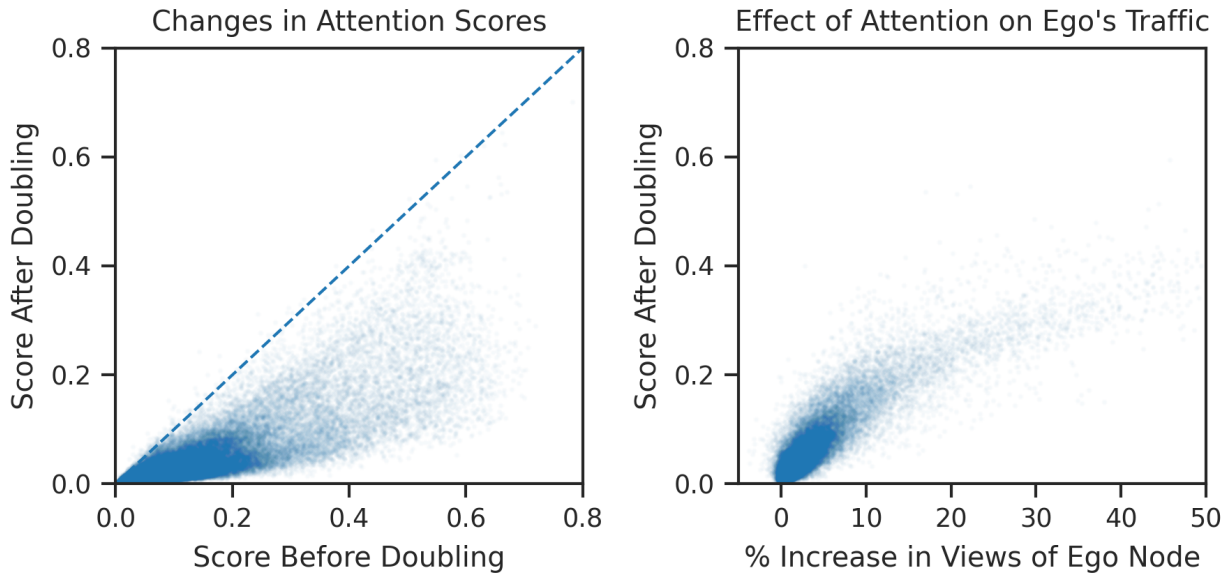


Figure 8: The effect of doubling a neighbor's view count during the forecast period. Left: scatter plot of attention scores before (x-axis) and after (y-axis) doubling. Right: Scatter plot of the relative increase of the ego node's views (x-axis) and attention scores after doubling (y-axis). Attention scores decrease as a neighbor becomes more popular (left). A higher attention corresponds to a larger flow of traffic from the neighbor to the ego node (right). See Section 2.4.5.

3 Measurement: Profiling global media audience during the COVID-19 pandemic

News media is an important political force, as it is sometimes referred to as the Fourth Estate[3] or fourth power, after the three separate branches – legislative, executive, and judicial – of a Western democratic political system. Conventional knowledge about this fourth estate has been challenged in recent years, by unexpected election results in different countries [20], by alleged interference from foreign powers faraway [8], by the widespread use of proprietary, algorithm-mediated content personalization [34], and by the fact that social media users influencing each other in real-time [5]. New understandings of how media reach and influence a connected audience base around the world are very much called for, especially since much research has focused on the United States (e.g., its media, users, and political systems) and a small number of Western countries.

This work aims to address three prominent gaps on measuring contemporary media. First, recent work on quantifying media ideological bias – via editorial curations [68, 2], surveys [21, 44], or data-driven measurements [10, 78] – all focus on estimating one single categorical label or numerical score for each media outlet, without explicitly taking into account the diversity of content from each outlet, nor the breadth of audience bases for each outlet. However, quantifying the distribution of audience reach (on their political leaning or other attributes) is important to reflect the diversity in both content providers and media consumers [59], hence the metaphorical "Shape of the Fourth Estate" in the paper title. The second is in measuring media reach beyond the US and several well-studied Western countries [63, 84, 22], and designing methods and metrics that can extend to a large part of the Global South. Quantitative profiles of media consumption in a diverse collection of countries could catalyze conceptual advances in characterizing media systems [47, 38], especially for societies outside of the Western democratic system and culture. The third is the analysis of cross-country differences in audience profiles. Existing studies on media either focus on in-depth analysis within a single country [10, 63, 78], or profile a number of countries separately by taking into account their respective political systems [50]. To the best of our knowledge, there have not been comprehensive (political) profiles of media audiences for global and country-specific outlets across different countries. Overall, our work seeks to answer two questions:

RQ1: For a given country, what is the breadth of readership, in terms of political leaning, for the top media outlets in that country?

RQ2: For a given media, how does its audience profile compare across different countries, and how does its audience profile compare with other media outlets in a given country?

To answer those questions, we start by collecting one of the largest COVID-19 tweet datasets from March to November 2020. We extract Twitter user locations based on tweet geotags and self-reported locations. We focus on the top 20 countries with the most number of geolocated users, yielding 17.7M Twitter users posting 346M tweets. We refer to this dataset as COVID2020 (Section 3.1). We further map the political leanings of global users onto one coherent spectrum, i.e., from lean-Dem to lean-Rep, by applying label propagation algorithm on user-user retweet networks (Section 3.2). We examine the overall user distributions in the top 20 countries of COVID2020,

[3] https://en.wikipedia.org/wiki/Fourth_Estate

and compare them to known theories in media systems (Section 3.3). We find that our estimates of average audience leaning score for each media are highly correlated with existing media bias ratings by conducting three different validations: (a) against six existing US-centric media bias estimates [2, 70, 10, 21, 78], for which we obtain similar correlation coefficients; (b) against the 2020 US election state-level voting results, for which we obtain correlation $r = 0.74$ with $p < 0.001$; and (c) against a recent media survey from 12 different countries [44], for which our estimates have at least $r = 0.67$ with 4 out of 6 intersecting countries.

We consider URL sharing behavior as a signal of media consumption, Our contribution to this work is the nuanced profiling of the audience leaning distribution for a specific media for users from a specific country (Section 3.4). In a country-centric view, we can show the country-level top consumed media and their audience leaning distributions. For instance, even though *thehill* has a Center bias rating, it is the second most lean-Dem media in Germany (more so than 5 other media rated Center-left). In a media-centric view, we can show the different audience bases of one media in different countries. For example, *nytimes* has a broader audience leaning distribution in the US and Brazil, but its audience distribution is much narrower in Germany and even more so in Australia. This means that the readers of *nytimes* in the US and Brazil would have access to a more diverse set of news sources, but not so for *nytimes*'s readers in Germany and Australia. We also find that sometimes a media's audience does not appear to align with its stated goals, or it can be more consumed by readers of the opposite partisanship in some countries. For example, *bbc*, the state-sponsored broadcaster in the UK, explicitly aims to reach a nationwide audience. But its audience base appears to be more lean-Dem than those of *telegraph* or *dailymail* in COVID2020. As another example, *breitbart*, considered a strongly lean-Rep outlet, has more than 75% lean-Dem audience in India and Pakistan, and between 25% and 50% of its audience being lean-Dem in Indonesia, France, Mexico, and Nigeria. Overall, our work enables a data-driven view of country-specific audience distributions, which provide new insights for media outlets, activist groups, and media watchdog organizations, as well as to individuals for reflecting on and choosing our day-to-day media diet.

The main contributions of this work include:

- A new, longitudinal Twitter dataset, COVID2020, with a high sampling rate and accurate user geolocation. The dataset will be released upon publication.
- A new set of media bias estimates (surrogated by the average user leaning score from users who share URLs from this media) that are shown to be highly correlated with existing editorial or survey-based scoring. This can be previewed from this anonymous link[4] and will be released publicly upon publication.
- Evidences that for the same media outlet, audience leaning distribution can differ significantly from country to country; within a country, relative audience leaning among outlets can differ significantly from known media labels.

3.1 COVID2020 Dataset

We collected public discourse about COVID-19 on Twitter from March 2020 to November 2020. Our dataset, referred to as COVID2020, consists of n=17708444 geolocated Twitter users from

[4] https://tinyurl.com/y53m8ttw

20 target countries and n=346304286 COVID-19 tweets posted by them. In this section, we first introduce our tweet collection strategy. We then describe the method of extracting user locations on Twitter. Next, we discuss the selection of political dividing hashtags. Lastly, we extract the URLs shared in tweets to study media consumption patterns about COVID-19 news.

3.1.1 Collecting COVID-19 tweets

We use the Twitter filtered streaming API to construct a new, high-volume COVID-19 tweet dataset. The tracked COVID-19 keywords are obtained from [24]. We ran the data crawler from March 2020 to November 2020, covering eight months during the first year of the pandemic. Our initial investigation in early 2020 shows that the daily volume of COVID-19 tweets exceeds the Twitter API limit by a large margin. For example, about 25M COVID-19 tweets were posted each day in March 2020. This is significantly over the Twitter API limit of 1% of total tweets, which is roughly 4M for a day.[5] If we just use the Twitter API naively, more than 80% of relevant tweets would be missing. A better data collection strategy is needed to increase the dataset coverage and recall.

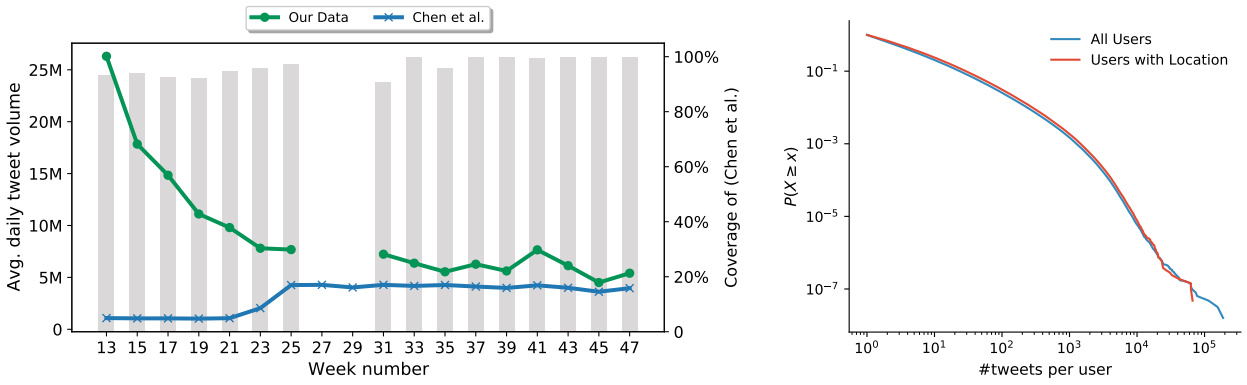
We adopt the crawling strategy proposed by [97], which consists of three steps: (a) partitioning the data stream into several sub-streams by tracked keyword and language; (b) estimating the sampling rate for each sub-stream using rate-limit messages provided by Twitter; (c) taking the union of all sub-streams. This strategy is shown to significantly reduce data loss, and thus results in a high-fidelity dataset for high-volume, real-time stream under Twitter API's limits. In our implementation, the estimated sampling rates range from 95% to 100% for different sub-streams, meaning that we have a high recall for all COVID-19 posts on the Twittersphere. To reduce the computational load, we processed one week's data every two weeks. The resulting dataset spans 18 calendar weeks (week 13 to week 47 in 2020). We experienced several server glitches during data collection, and lost data for two entire weeks (week 27, 29) and another four days (in week 17, 31). In total, we obtained n=999040035 COVID-19 tweets posted by n=62687121 users.

The temporal data volume of COVID2020 is shown in Figure 9a (green line). The x-axis is the week number. The left y-axis shows the average number of daily tweets in each selected week. The volume of COVID2020 dataset starts from more than 25M tweets per day in March, and drops to about 5M tweets per day in November. The decline may be attributed to user fatigue on COVID-19 topics. According to a survey from Pew Research Center [72], 71% of Americans expressed willingness to take breaks from COVID-19 news, and 43% said the news made them emotionally worse.

We also compare our COVID2020 dataset with a widely cited COVID-19 tweet dataset curated by [24] (Figure 9a blue line). We find that COVID2020 is significantly larger throughout the data collection period, though the volume difference decreases as time passes. For example, in week 13 (03/23 - 03/29), [24] had only 7.56M tweets while we collected 184M tweets (24 times larger). Note that the observation of decreasing volume in COVID-19 tweets is only made possible by our data stream with a high-sampling rate (green line). On the contrary, simply reading daily volumes from a single Twitter stream would lead to the opposite conclusion (blue line).

On the right y-axis, we show the fraction of tweets from [24] that also appear in COVID2020 for every week (gray bar). On average, 96.59% of tweets from [24] are included in COVID2020. To

[5] Researchers have found that Twitter streaming API provides no more than 50 tweets per second [54], though the Twitter API limit is more commonly known as no more than 1% of the entire tweet volume for any time intervals.



(a) (b)

Figure 9: (a) Comparison between COVID2020 dataset and [24]. Left y-axis: our COVID2020 (green line) is much larger than [24] (blue line) throughout time. Right y-axis: most tweets (on average 96.59%) from [24] are also included in COVID2020 (gray bar). (b) CCDF plots of the number of tweets per user for all users (blue line) and the set of geolocated users (red line).

the best of our knowledge, COVID2020 is the largest COVID-19 tweet dataset for the same period.

3.1.2 Identifying Geolocated Twitter Users

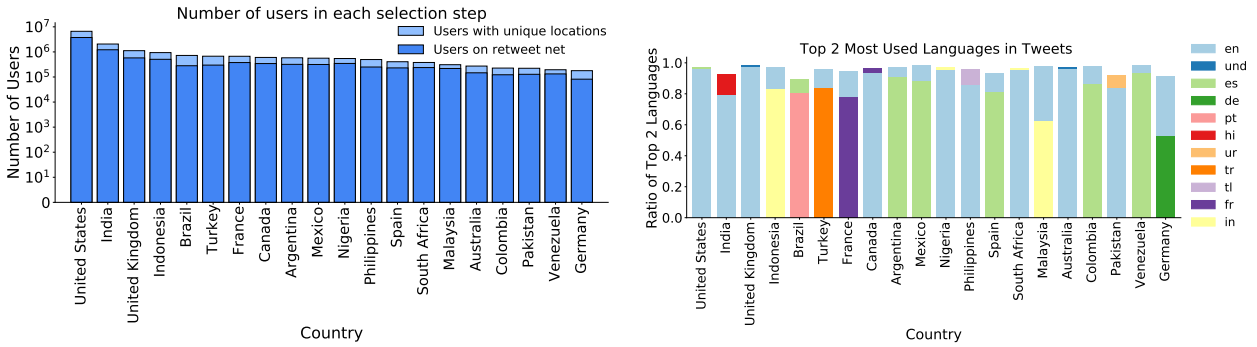
To measure the user behavior in different countries, we extract the location information from (a) Twitter users who use geotags in their tweets and (b) those who explicitly mention their locations on their profiles.

- **Geotagged users.** Approximately 0.85% of all tweets in COVID2020 are attached with geotags. A geotagged tweet includes a (city, state, country) place tag and may also include a (longitude, latitude) coordinates tag. We use the country tag to identify users worldwide, and further use the state tag to map US users into the 50 US states. To aggregate geotagged tweets into geotagged users, we assign the identified locations to users, which essentially builds the user mobility traces during COVID-19. For simplicity, we filter out all users posting geotagged tweets from multiple countries and US users from multiple US states. In the end, we identified n=1958826 geotagged users.

- **Geoparsed users.** Because most users do not post geotagged tweets, we also parse the self-reported location strings from their user profiles. We use the *Simplemaps*[6] data to construct a list of all possible combinations of cities, states, and countries (both full names and abbreviations). We then check the location strings against the curated list. We only include users with one exact string-matching location. We identified n=20975218 geoparsed users.

At the country level, there are n=1077887 overlapping users between the two aforementioned sets. We consider the locations obtained by geotagged tweets as ground-truth and then evaluate the precision of string-matching geoparsing on overlapping users. We find that geoparsing is able

[6] https://simplemaps.com/data/world-cities



(a) (b)

Figure 10: (a) The size of users with unique locations (lighter blue bars) and users who are retweeting others (darker blue bars). Countries are ranked by the number of users with locations. (b) The ratio of the top two most used languages in 20 selected countries.

to identify the locations of 93% of overlapping users correctly. At the US state level, there are n=285808 overlapping US users. Geoparsing correctly identifies the state locations of 92% of them. The high accuracy shows that it is possible to take advantage of both methods to extract user locations on Twitter. We hence merge the two sets of users, with mismatched geoparsed users assigned locations from geotagging. In total, there are n=21097109 users with one unique geolocation. They posted n=397943516 tweets.

Figure 9b plots the complementary cumulative density function (CCDF) of the number of tweets per user. For all users in COVID2020 and for the set of geolocated users, the two distributions are very similar. This observation assures us that analyzing geolocated users would not skew the user activity distribution, either to the more active or less active subpopulations.

There are several third-party libraries for extracting geolocations from social media data [69], for example, OpenStreetMap and Google Geocoder API. We experimented with those libraries in our pilot test. However, we find that those libraries often operate based on a quota system and thus are unable to process millions of requests in a reasonable amount of time. We also evaluated several machine-learning-based and rule-based geoparsing libraries (e.g., Mordecai) that have been trained on location texts, but found their performances not satisfactory with the noisy location strings. Therefore, we use the geotags in tweets and exact string-matching for self-reported location on user profile.

Geolocated users broken down by country. In this work, we focus on the top 20 countries with the most geolocated users. Figure 10a shows the size of geolocated users in the top 20 countries in shallow blue bars. US has the most users (n=6638437) while Germany has the least (n=178571). Note that the ratio of geolocated users across countries is on par with previous studies, for example, the ratio of US, Spanish, and German Twitter users reported by [11]. For each country, we also plot the size of users who have retweeted others in darker blue bars. We will infer the political leaning of those retweeting users in Section 3.2.

Top two tweet languages in each country. We also want to find out whether the COVID2020 dataset is biased toward English tweets because our tracked COVID-19 keywords are all in English.

To address this concern, we count the frequency of used tweet languages in the top 20 countries. The language information is extracted from the "`lang`" field in the tweet object. It can be one of the BCP 47 language codes or "`und`" if the language is not recognized.

Figure 10b shows the stacked bar plot of the top two most used languages in each country. The y-axis shows the ratio of tweets written in marked language. We find that a large fraction of tweets are posted using native or official languages in non-English speaking countries. For instance, Spanish (`es`) is the most used language in countries like Argentina, Mexico, Spain, Colombia and Venezuela; Portuguese (`pt`) is most used in Brazil; French (`fr`) in France; and Turkish (`tr`) in Turkey. This assures that the COVID2020 dataset covers Twitter users who are representative of their resident countries, and thus not a biased group of English-speaking users in all countries.

3.1.3 Curating Political Hashtags Related to 2020 US Presidential Election

We initially wanted to use one universal scale to measure political polarization all over the world, but we find that it is an intrinsically difficult task due to the multifaceted nature of political systems operating in different countries. The two widely used political scales by computational social scientists are the left–right and liberal-conservative spectrum. However, neither of them can be applied unanimously worldwide without careful calibrations. Firstly, the views about contested political topics are misaligned across countries. For example, US liberals support abortion while conservatives oppose; however, both UK liberals and conservatives support abortion [76]. If we use pro/anti-abortion keywords to classify users, we would misclassify UK conservatives as liberals. Secondly, the political dividing topics vary significantly across countries. For example, in the US, those topics include abortion, gun control, racial justice, etc [61]. But in most European countries, debates between political groups focus on taxation, welfare, and immigration [22]. It is even unclear what topics are debated in traditionally less studied Asian and African countries. One strategy that might address this problem is to curate a specific list of political dividing topics for every single country. However, it is difficult without the help of political scientists who have domain knowledge about individual countries.

In this work, we focus on a narrower perspective of the political spectrum for the global Twitter users. We measure the leaning toward the two major US political parties, i.e., the Democratic party and the Republican party, in the 20 selected countries (lean-Dem and lean-Rep for simplicity). Because our COVID-19 data collection period is leading up to the 2020 US presidential election, many tweets in COVID2020 also contain hashtags related to the US election. We collect political hashtags curated in previous work [52, 94], for example, *#bidenharris2020*, *#voteblue* for lean-Dem hashtags; *#maga*, *#trump2020* for lean-Rep hashtags.

In total, we obtained 414 lean-Dem and 409 lean-Rep hashtags Across all countries, there are n=331233 users using these hashtags on their Twitter profile descriptions and/or in any posted tweets. n=271967 (82.1%) of them are located in US. They are referred to as *seed users*, and are later used to infer political labels of a larger user population in Section 3.2. Figure 11 displays the top 25 lean-Dem and lean-Rep hashtags, ranked by the number of users who have used them. Darker bars in the front indicate seed users in the US while lighter bars in the back indicate seed users across the world. Though the top ranked hashtags have been mainly used by US seed users, there are a few tags that have a large presence outside of the US. For instance, *#covididiots*, the most used lean-Dem hashtag, was used by n=12602 users in the US, and n=31624 users in all other countries.

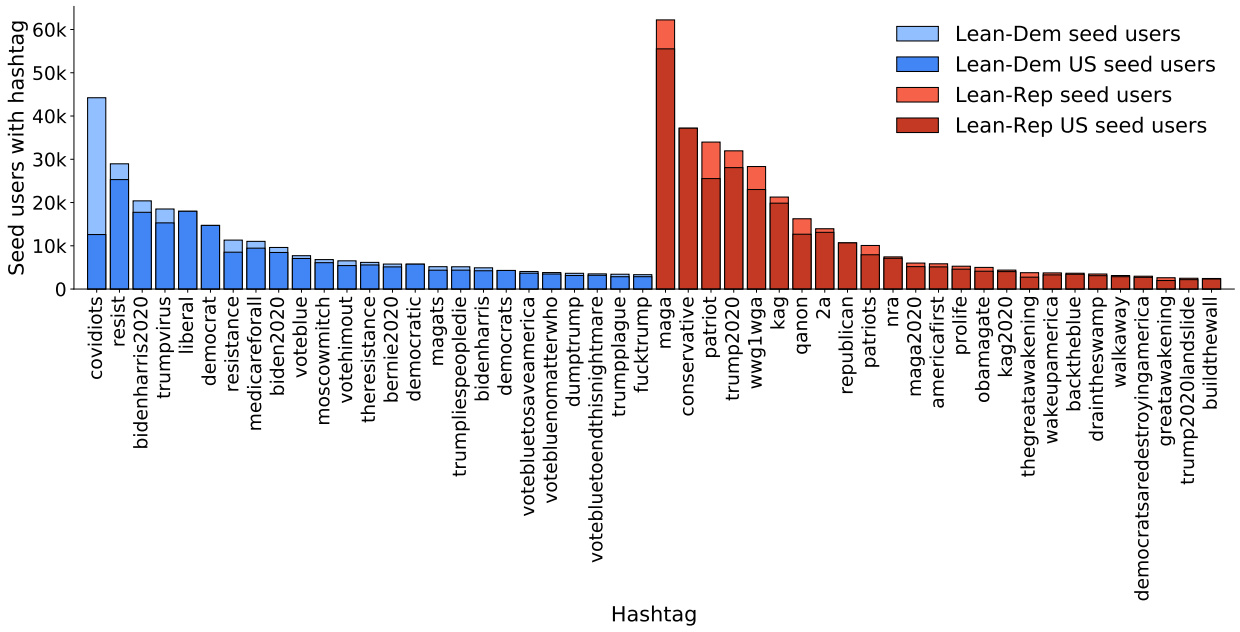


Figure 11: Top 25 lean-Dem and lean-Rep hashtags ranked by the number of seed users using the corresponding hashtag.

3.1.4 Extracting URL Domains

URL sharing behavior is commonly used as a proxy for studying media consumption patterns online [3, 10]. To understand how Twitter users in different countries consume and share COVID-19 news, we extract URLs from each tweet in COVID2020. Specifically, we use the `expanded_url` field in the tweet object. Twitter API provides the full URL address if the URL is shortened by Twitter's own shortener (i.e., *t.co*). We further curate a list of media-specific shorteners (e.g., *cnb.cx* to *CNBC*, *wapo.st* to *Washington Post*) and use them to resolve shortened URLs. We discard URLs that are shortened by other shorteners (e.g., *bit.ly*, *tinyurl.com*) because expanding such shortened URLs often requires sending an http request to the browser, which is very time-consuming at scale. For each URL, we extract the domain information. There are n=28497 domains shared by 50 or more users from the top 20 countries in COVID2020 dataset, and n=8410 domains shared by more than 50 users in the US.

There are two desired benefits of studying media consumption at the level of media domain. First, even a URL has only been shared by a handful of users, aggregating into the domain level would give us more users to analyze. Second, for the same media, different domains may reveal the coverage focus. For example, one would expect that *cnnphilippines* and *cnnindonesia* have very different audience bases because the former focuses on reporting Philippines news while the latter focuses on Indonesia, even though they both belong to the CNN media group.

For each domain, we define a metric *global audience reach*, computed as the number of unique Twitter users sharing URLs from this domain across all top 20 countries. Figure 12 displays the top 50 domains according to their global audience reach. We can see many established media from English speaking countries spanning the full range of the political spectrum (e.g., from *cnn* in US and *bbc* in UK, to *foxnews* in US and *independent.co.uk* in UK). Media outlets from Asia also appear in the top 50, such as *cnnindonesia* in Indonesia, *astroawani* in Malaysia, *aajtak.in* in India.

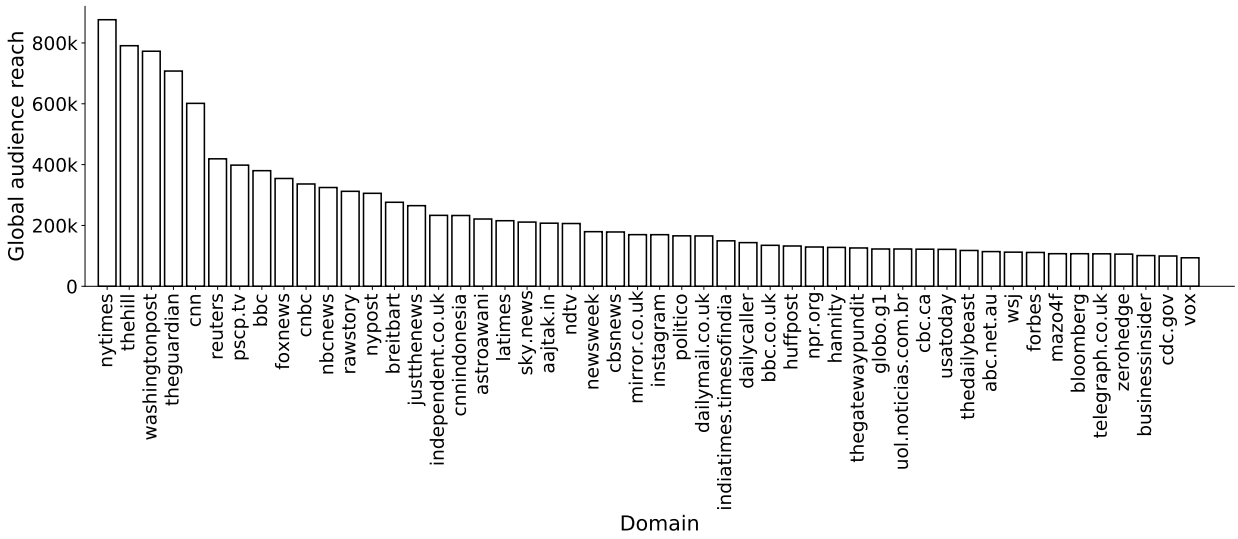


Figure 12: Top 50 domains ranked by their global audience reach.

Despite using the word "media", we notice that some media platforms also appear in top positions (e.g., *pscp.tv*, *instagram*). They are websites that host content from others rather than producing content by themselves. We opt to keep those platform domains, since many media outlets also function as a platform for user-generated content, and drawing a precise line will be difficult.

3.2 Mapping the political leaning of global users onto one scale

We develop a procedure for estimating the political leaning toward the two major US political parties, i.e., lean-Dem or lean-Rep. The procedure contains four steps:

1. **Identifying seed users.** We use hashtags on both the users' profiles and their tweets to label the leaning of a set of seed users. The process of curating political hashtags related to 2020 US presidential election is described in Section 3.1.3. Next, we calculate a leaning score for each user by $\frac{rep-dem}{rep+dem}$, where rep and dem are the numbers of lean-Rep and lean-Dem hashtags, respectively. To reduce noises, we only consider users with a score higher than 0.9 (less than -0.9) as lean-Rep (lean-Dem) seed users. There are n=331233 seed users (n=177170 lean-Dem, n=154063 lean-Rep) in total.

2. **Constructing user-user retweet network.** A simple retweet (without comment) disseminates the original post to the retweeter's followers, showing support from retweeter to retweetee. Researchers have shown that the retweet network is of high homophilyand can be used to estimate user leaning on Twitter.To this end, we construct the user-user retweet network for the geolocated users in the 20 target countries. The result is an undirected network. Each node is a user, and each edge indicates at least one retweet between the two users. We use the number of retweets as the edge weight. The number of users on the retweet network within each country is shown by the darker blue bars in Figure 10a.

3. **Extracting network backbone.** To remove insignificant edges in the network, we run the disparity filtering algorithm [80] to extract the network backbone. We set the significance threshold

Political leaning	Precision	Recall	F1 score	#Identified users
lean-Dem	0.86	0.94	0.90	n=1850270
lean-Rep	0.92	0.82	0.87	n=589601

Table 6: Performance of estimating user political leaning via label propagation in 10-fold cross validation setup.

to be 0.01. We make a minor modification to the filtering process: if an edge is statistically insignificant but connects nodes that are both seed users, we do not discard it to ensure better connectivity to seed users. This modification is shown to help the label propagation step described later. The reduced network has n=2439871 nodes with n=15820431 edges in total.

4. **Propagating labels from seed users.** Finally, we apply the label propagation algorithm [102] on the extracted backbone network to infer user leanings. An important hyperparameter is α, which controls the trade-off between preserving original information and receiving new information from neighbors. We used 10-fold cross-validation to do a line search between 0 to 1 and found the optimal α to be 0.85. Upon convergence, label propagation returns a score for each leaning between 0 and 1, which is then re-scaled into $[-1, 1]$. We use the normalized score as the predicted leaning score for each user. A score closer to -1 (1) means this user's retweet network mainly consists of lean-Dem (lean-Rep) users. A score closer to 0 means this user's neighboring network consists of a mix of lean-Dem and lean-Rep users. We also use 10-fold cross-validation to estimate the performance of label propagation. Table 6 shows that our estimation has high F1 scores for both lean-Dem and lean-Rep users. Note that we estimated political leaning for n=2439871 users. This is 7.4 times of the seed users, and makes up 11.6% of all geolocated users.

Our proposed procedure has one desired feature for cross-country analysis: it calibrates and unifies the scales – which are often perceived differently – across countries. For example, US conservatives and UK conservatives have contrasting opinions on issues such as universal healthcare, LGBT rights, and gun control. In this work, we choose to map the political spectrum onto the axis of the US Democratic-Republican scale. This makes our estimated quantity comparable across different countries.

3.3 Distribution of political leaning across countries

We profile the distribution of uses' political leaning in each country. Figure 13 displays 20 density plots for the estimated user leanings, one for each of the top 20 countries in Figure 10a. Leaning scores in $[-1, 0)$ are considered lean-Dem and colored in blue, while scores in $(0, 1]$ are considered lean-Rep and colored in red. Note that for each country, the number of users with leaning estimates is smaller than the total number of geolocated users in Figure 10a.

Our profiling shows that lean-Dem users are the majority in all of these 20 countries (ranging from 50.7% to 94.6%). This is consistent with a COVID-19 era survey by Pew Research Center that 69% of top US Twitter users are considered lean-Dem [81]. Another overall impression is the considerable diversity of user distributions across different countries. While countries such as the US and Canada exhibit a bi-modal distribution with one mode (local maximum of a distribution)

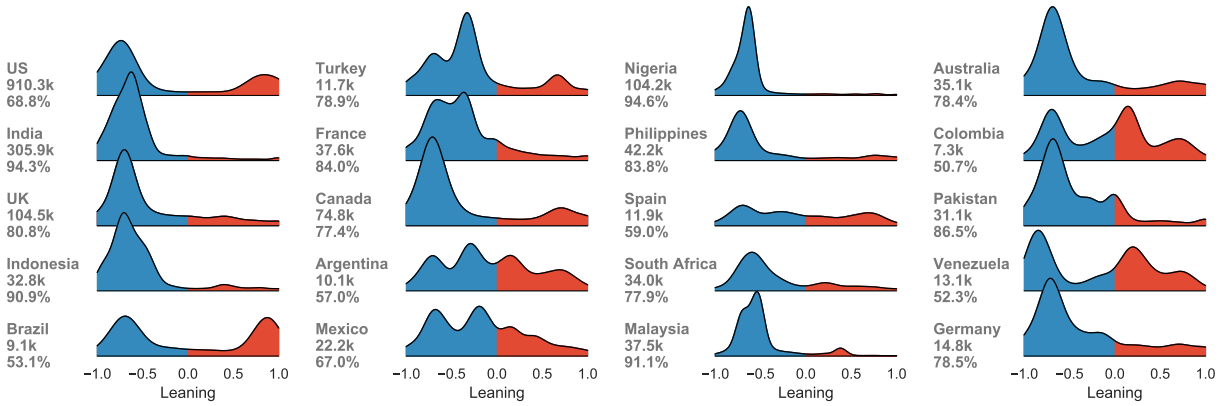


Figure 13: The distributions of estimated user leaning in 20 countries from COVID2020. x-axis: leaning scores obtained from label propagation; y-axis: relative frequencies (density) of users within a country. The number below country names are the total number of users found in this country from COVID2020, and the fraction of lean-Dem users in the country.

in each of the lean-Dem and lean-Rep regimes, user distributions in many countries show multiple modes, such as the two groups of lean-Dem users in France, as well as the multiple modes in both the lean-Dem and lean-Rep regimes in Argentina and Mexico. In a number of countries, there is a mode near or at the center (such as Colombia and Pakistan), showing a wide variety in the division of political leaning that does not align with the US-centric ideological manifests.

How shall we conceptualize these observations? Posts on Twitter are known to behave more like broadcast media than a social network [60]. We posit that comparative studies on media systems in different countries will help explain and conceptualize our observations. In the three original models of Western media systems [47], US, UK and Canada are classified as the *liberal* system. The user distributions in these three countries indeed exhibit high similarity – large spread over the lean-Dem to lean-Rep spectrum, and having a mode in each of the lean-Dem and lean-Rep regime. There are also deviations from existing media system theories, for example, France and Spain are both classified as *polarized pluralist*, but Twitter users in France appear to have two prominent groups in the lean-Dem regime, whereas users in Spain are more evenly distributed. We acknowledge that the top 20 countries are merely a sample of all possible analyses that one can do to compare country-level political profiles, comparisons of *democratic corporatist* will be possible if one examines the top 50 countries in **??** in **??**. The top 20 countries in COVID2020 provide a number of observations on countries outside of Western Europe and North America. Since [47]'s work, there has been a proliferation of typologies. One overview is given by [38]. There are a large number of factors being identified as relevant; they include but are not limited to communication values (freedom *vs.* responsibility), government system, media ownership and interaction with the state, geopolitical history, and philosophy. Those factors often correlate with location and the main language of the country. There are a few groups of countries that support such multifaceted similarities, such as (a) Argentina and Mexico, Indonesia, Philippines and Malaysia; (b) Colombia and Venezuela. Some countries may violate surface similarities, such as users from India are heavily lean-Dem, whereas users in Pakistan have one lean-Dem mode and one center-leaning mode.

What can or cannot be read from these plots? The user leanings are estimated by whom are connected to those using lean-Dem and lean-Rep hashtags via retweets. While we do not claim that this user population is representative of the country or even its Twitter user base, in Section 3.4 we show that media leanings from this dataset have strong correlations with those obtained via traditional survey methods [44] in Australia, US, UK, Spain, but weaker correlation for France and Germany. This means that the interpretability of the political profile results here varies by country. Overall, the geolocated users in COVID2020 dataset provide a unique and geographically diverse picture of user polarization around the world, and we hope it will shed light on understanding political behavior and media landscape in countries that do not often appear in computational social science writings, such as Malaysia, South Africa, Nigeria. We hope our methodology and initial observations serve as a basis for further inquiry.

The observations in this section focus on all users and all media domains within a country, in Section 3.4 we will break down the user distribution for each media domain.

3.4 Media consumption patterns across countries

From the sharing behavior of geolocated users with political leaning estimates, we obtain the distribution of audience leanings for each media domain for a given country or US state (Section 3.4.1). We then validate the domain leaning scores against known media rating sources and recent research results, we also examine the correlation of domain leaning to voting records in US (Section 3.4.2). Lastly, we profile the user leaning distributions for media domains across different countries (Section 3.4.3).

3.4.1 Computing Audience Leaning Distributions for Each Domain

In this work, each Twitter user can be represented as a nested tuple $(u, l_u, \nu_u, \mathcal{D}_u)$, where u is the user id, l_u is the location of user u (Section 3.1.2), ν_u is their political leaning (Section 3.1.3 and Section 3.2), and $\mathcal{D}_u = \{d_{u1}, d_{u2} \ldots\}$ is the set of media that the user has shared (Section 3.1.4).

We define the total *audience reach* $\kappa(d)$ of domain d as the number of unique users sharing URLs from domain d: $\kappa(d) = |\{u | d \in \mathcal{D}_u\}|$. Similarly, the *audience reach* of domain d in location l is $\kappa(d, l) = |\{u | d \in \mathcal{D}_u, l_u = l\}|$, where $|\cdot|$ denotes set cardinality.

We compute the *average audience leaning* $\bar{\nu}(d)$ of each domain d by averaging over all users who have shared it at least once. We can similarly compute the average leaning of domain d shared by users in location l, denoted as $\bar{\nu}(d, l)$. In subsequent uses of the average audience leaning score, we omit the range of domains d and (a given) location l when they are clear from context.

$$\bar{\nu}(d) = \frac{1}{\kappa(d)} \sum_{\{u | d \in \mathcal{D}_u\}} \nu_u; \qquad \bar{\nu}(d, l) = \frac{1}{\kappa(d, l)} \sum_{\{u | d \in \mathcal{D}_u, l_u = l\}} \nu_u \ .$$

Moreover, information extracted from the COVID2020 dataset affords a richer set of statistics for each domain. We denote the distribution of user leanings for domain d, in location l as $\text{Distr}_\nu(d, l)$, which can be represented as mean and variance, median and quantile values (Figure 18 and **??**), or density plots (Figure 13). Note that similar user-based aggregation was used in recent work: [10] described a domain by averaging over the leanings of users who had shared it; [78] computed domain scores based on Twitter users who had shared its URLs.

3.4.2 Validating Average Audience Leanings

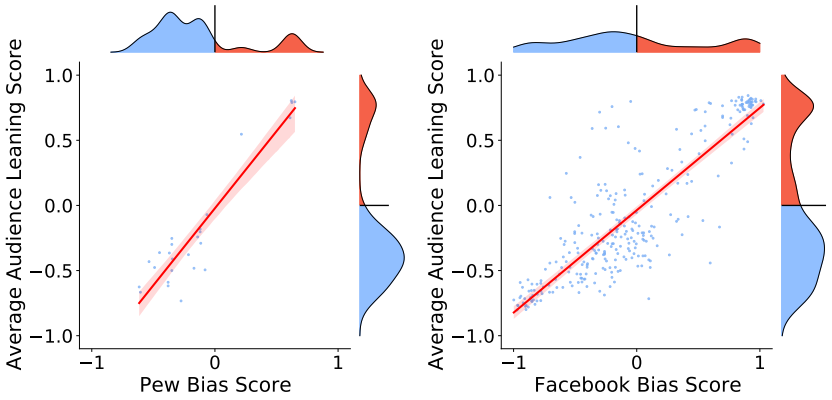
Comparing with existing media bias ratings for US audience. We first compare the domain leaning scores to six other estimates in the recent literature [10, 21, 70, 78, 2]. Since all of these sources are based on the US and report a numerical average or categorical label, we compare them to the average domain leaning for US users $\bar{\nu}(d, l = \text{'US'})$. We include domains with an audience reach (unique number of users) of at least 50, there are 8,410 such domains in the US portion of COVID2020. We first take the intersection between these domains and those from each of the prior literature; we then compute Pearson's r for continuous scores and Spearman's ρ for ordinal scores, between average domain leanings $\bar{\nu}$ and the scores from existing sources. Section 3.4.2 summarizes the results with scatter plots of the scores and correlation results.

Pew Research audience profile score [70] was collected from *n*=2901 web respondents that were representative of the US Internet users in 2014. We reconstructed these scores from the interactive webpage [71]. The remaining five sets of media bias scores are obtained from [78, 10, 21, 2]. Figures 14a to 14c show high correlations between $\bar{\nu}$ and survey-based [70], audience-based [10] and crowd-sourced [21] audience scores. The correlation of $\bar{\nu}$ with the AllSides editorial scores is fairly high ($\rho = 0.76^{***}$). In particular, the correlation of $\bar{\nu}$ with the Pew audience scores [70] and AllSides score [2] (curated in 2018 by [78]) are higher ($r = 0.93^{***}$ and $\rho = 0.76^{***}$ respectively) than those from [78]'s Twitter-sharing based scores ($r = 0.78^{***}$ and $\rho = 0.64^{***}$ respectively). The lowest correlation is observed with the Twitter-based scoring and MTurk scoring, both obtained from [78]. The coverage in all domains from [78] is also low (18%) due to keyword filtering that led to COVID2020 narrowing down the candidate set of domains (as intended). The correlation with MTurk rating scores is lower ($r = 0.48^{***}$), but it is consistent with [78]'s own observation($r = 0.50^{***}$).

Comparing with US state voting results. Figure 15 shows the regression plot between the mean bias scores of URLs shared by all users within each US state and the percentage of votes to the Republican party in 2020 US presidential election. We observe a strong correlation ($r = 0.74^{***}$, $p < 0.001$) between our estimated domain bias scores and state-level voting results.

Comparing with international media surveys. A key contribution of our study is the cross-country analysis. To this end, we compare $\bar{\nu}$ with survey results conducted by [44], in which respondents form a stratified sample of subjects from 12 different countries, and were asked about their political leaning on a seven-point scale (ranging from -0.5 to 0.5). Respondents were also asked about the news outlets they had read online and offline in the past weeks, from a candidate list of 30 popular outlets that varied from country to country. [44]'s audience political leaning score for each news outlet is the mean of the self-identified leaning scores of its audience. We find a strong correlation between the media bias scores from our estimation and that from [44] in US and Australia ($r > 0.8^{***}$). We also notice a moderate correlation in Spain and UK ($r \geq 0.7^{***}$). The correlation with France and Germany are considerably lower ($r = 0.45$ and 0.27, respectively) and not statistically significant. This is likely due to the narrow leaning distribution of $\bar{\nu}$ in France, and that survey outcomes for German outlets are all around the center.

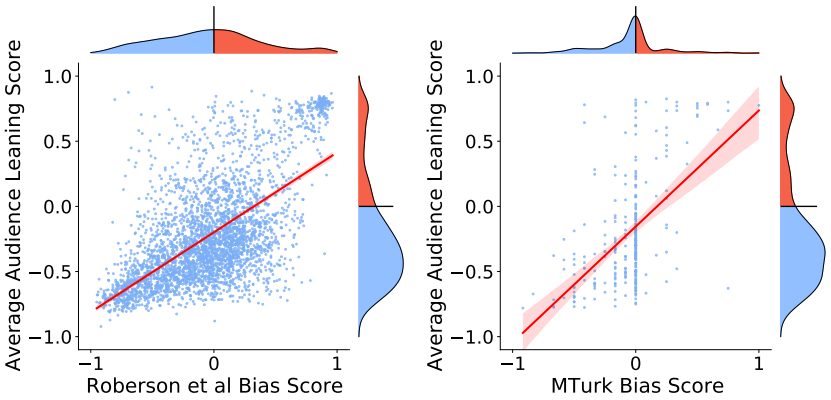
Across all three validation tasks, we observe strong correlations between $\bar{\nu}$ and other estimates of US domain biases, voting records, and international media leaning surveys. This provides us great confidence in producing new observations on the media sharing behavior internationally,



(a) [70], n=26 (93%), $r = 0.93^{***}$ (b) [10], n=299 (69%). $r = 0.86^{***}$

(c) [21], n=15 (100%). $r = 0.83^{***}$

(d) [2], n=152 (78%), $\rho = 0.76^{***}$



(e) [78], n=3517 (18%), $r = 0.63^{***}$

(f) [78], n=281 (73%), $\rho = 0.48^{***}$

Figure 14: Correlation between average audience leaning from our data versus bias scores from other sources, each with "[reference], n=number of overlapped domains (fraction covered in the other source), correlation score" – Pearson's r for continuous scores and Spearman's ρ for ordinal scores. *** denotes $p < 0.001$. Ordered chronologically (oldest to newest) from (a)–(f). Side densities represent the distribution of domain leaning scores from either source.

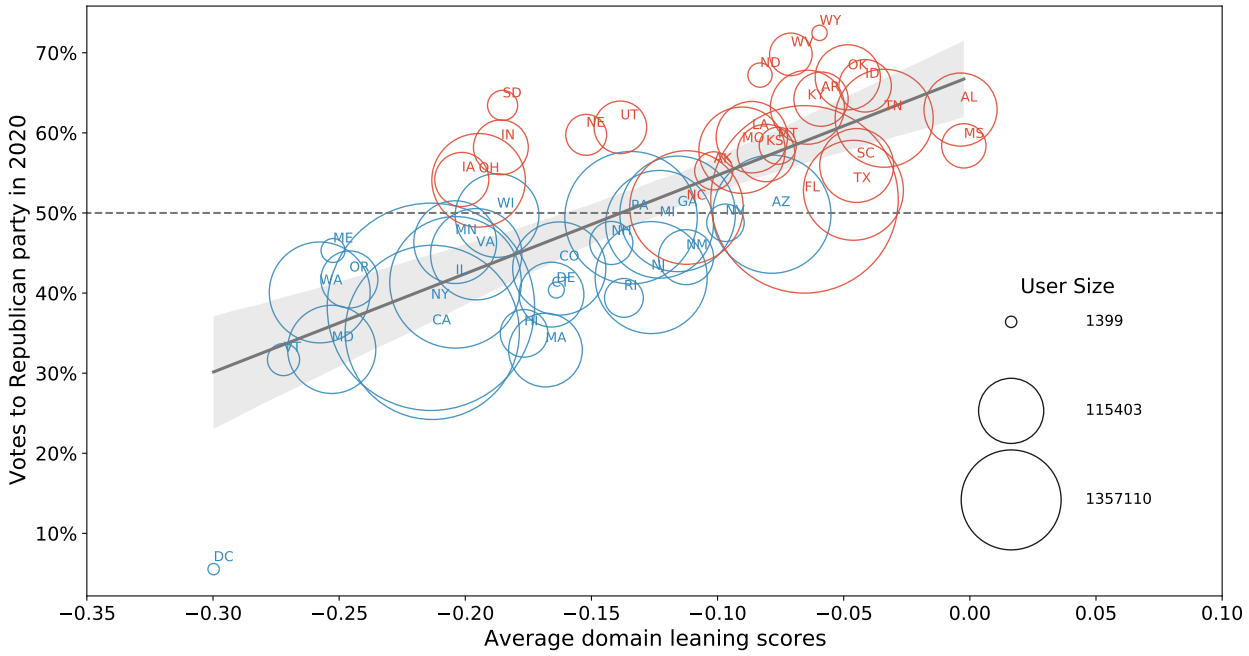


Figure 15: Regression plot between the average domain leaning scores and proportion of votes to the Republican candidate in the 2020 US presidential election in each US state. x-axis: average leaning scores of URLs shared; y-axis: percentage of votes to the Republican party. Red color indicates more than 50% votes to the Republican party while blue indicates more than 50% more votes to the Democratic party. Size of the bubble are proportional to the number of geolocated Twitter users in each state. Pearson's $r = 0.74^{***}$; $p < 0.001$.

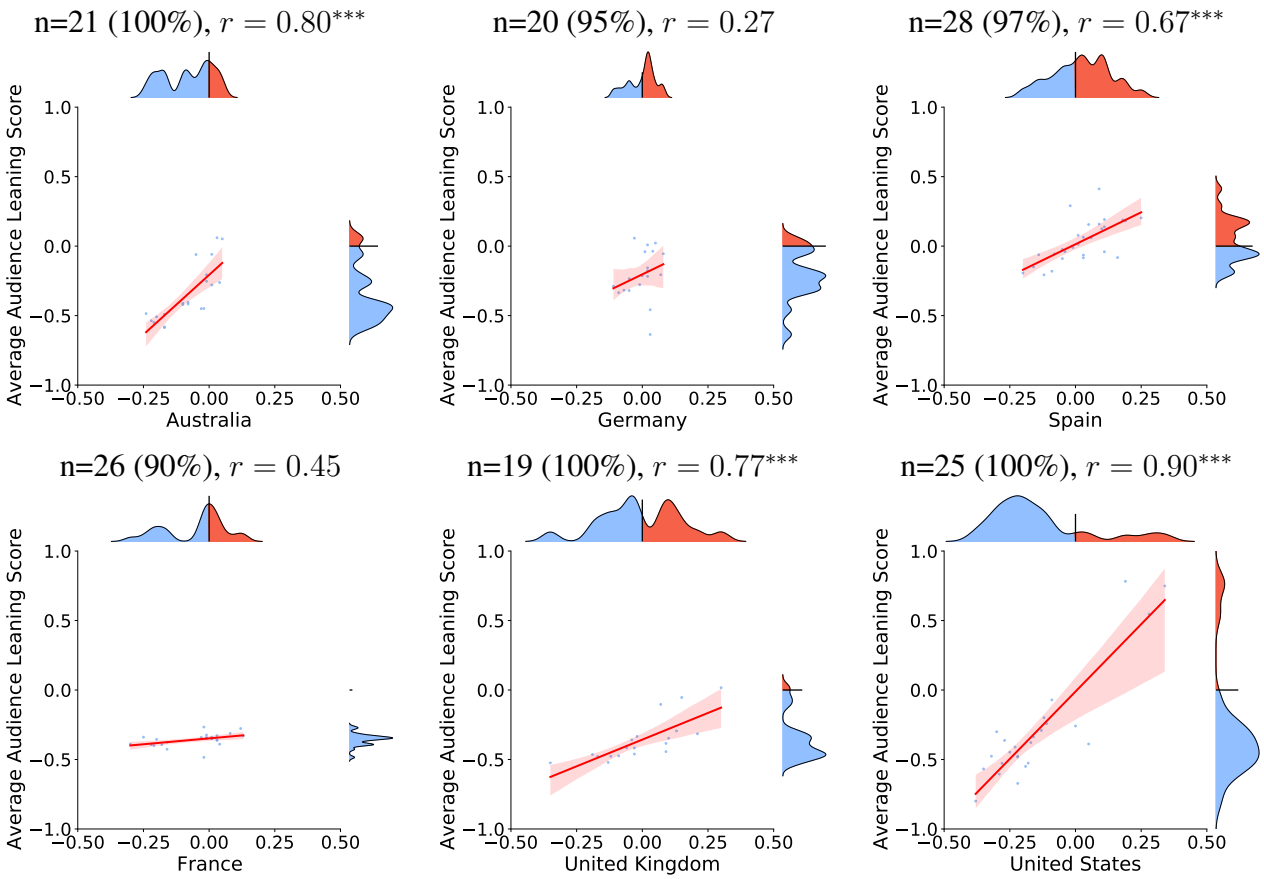


Figure 16: Correlation between our average audience leaning score(y-axis) versus those from [44] (x-axis). We use Pearson's r correlation test. $^{***}P < 0.001$. *n=xx* shows the number of overlapped domain with [44], *(xx%)* is the coverage of all domains from [44]. For domains shared in multiple countries, we calculate the average leaning scores in each country independently.

though we caution that the reliability of $\bar{\nu}$ in countries other than those validated above needs further scrutiny.

3.4.3 Profiling Global Media Consumption

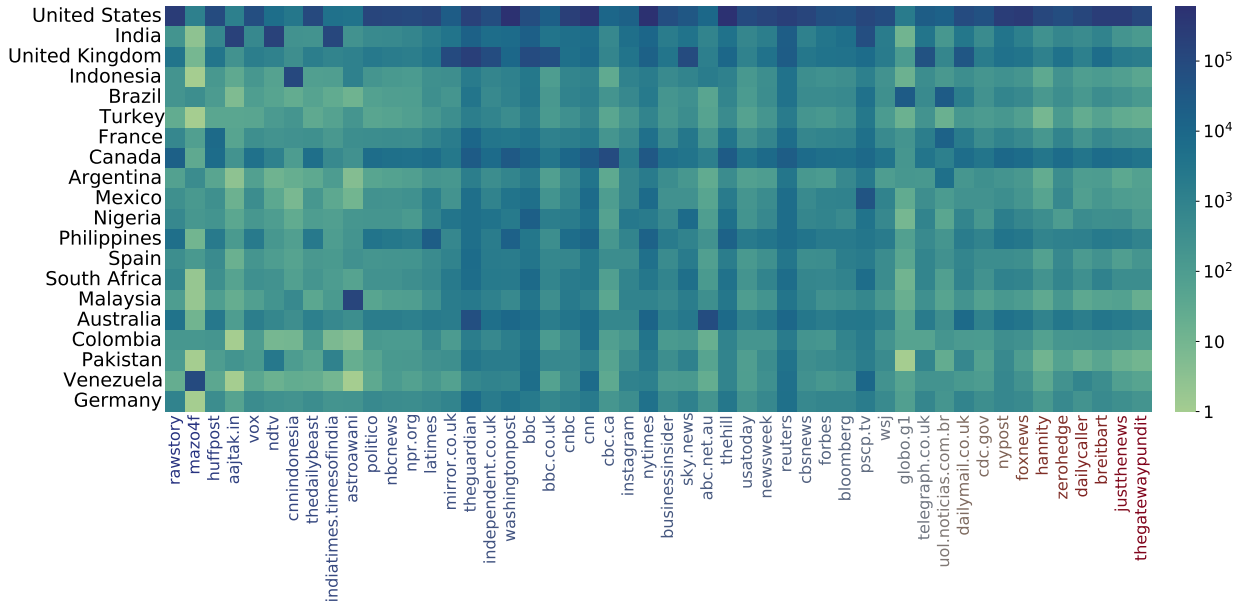


Figure 17: Heatmap of audience reach $\kappa(d, l)$ of the overall top 50 domains in the top 20 countries of COVID2020. x-axis: domain names ranked by their average audience leaning score over 20 countries, colored according to $\bar{\nu}$ from lean-Dem (left) to lean-Rep (right). y-axis: 20 countries ordered by the number of users. Cell color represents $\kappa(d, l)$ for country l and domain d, in log scale.

Figure 17 contains a headmap of audience reach of the top 50 media domains (extracted in Section 3.1.4, ordered by average audience leaning) across the top 20 countries (extracted in Section 3.1.2, ordered by number of users). In this map, US has the most number of users sharing a large subset of the top 50 media. We can also identify a dozen of media domains for which the US is not the country with the most audience. These include country-specific outlets such as Aaj Tak (*aajtak.in*) in India, The Mirror (*mirror.co.uk*) in UK, CNN Indonesia (*cnnindonesia*) in Indonesia, G1 (*globo.g1*) in Brazil, CBC News (*cbc.ca*) in Canada, Astro Awani (*astroawani*) in Malaysia, ABC News (*abc.net.au*) in Australia, and Con El Mazo Dando (*mazo4f*) in Venezuela. There are a few other domains with significant reach in more than one non-US country, such as The Guardian (*theguardian*) with significant audience in UK, US, Australia and Canada, and UOL (*uol.com.br*) with significant audience in Brazil, Argentina and France. A few countries among the top 20 (e.g., Germany, France, and Turkey) do not show a significant user base among the top 50 media domains – while we choose to present 50 domains due to readability, one would expect to expand the domain list for a comprehensive analysis in these countries (see Figure 18).

In the rest of this section, we examine this media profile centered on countries and media domains, respectively.

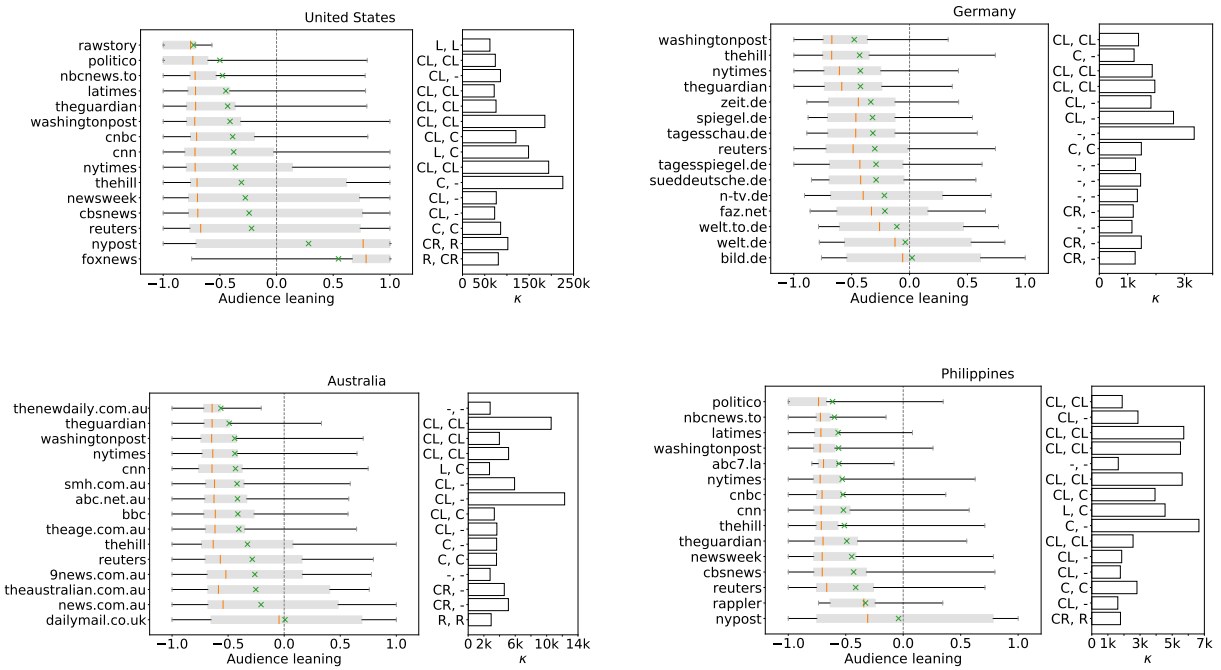


Figure 18: Audience leaning distribution for the top 15 media domains ranked by audience reach in four countries. Left subfigure: box plots on audience leaning distribution. x-axis: audience leaning; y-axis: domain names. Green cross: mean; orange line: median. Whiskers are from the 10^{th} percentile to the 90^{th} percentile. Right subfigure: bar plots on audience reach κ of each domain and country. x-axis: audience reach in thousands. y-axis label: domain ratings from [68] and [2], "-" means not rated by the corresponding source. The domains are ranked by average audience leaning $\bar{\nu}$ (green cross).

A country-centric view. Figure 18 profiles the distribution of audience leaning scores, for the top 15 media domains (by audience reach) in four example countries: the US, Australia, Germany, and Philippines. Figures on all of the top 20 countries are in **??**. The media domains within each country are ordered by their average leaning score $\bar{\nu}(d, l)$ (shown as a green cross), we also include media labels from AllSides [2] and Media Bias Fact Check (MBFC) [68] where available. Both labels are presented as 5 categories (L - Left, Center Left / Lean Left - CL, Center - C, Center Right / Lean Right - CR, R - Right). Media in MBFC categories "Extreme Left/Right" do not appear in the top 15 domains. Comparing the two editorially curated ratings, MBFC covers more news outlets (especially outside of the US), and AllSides uses different sources of information, including editor ranking, staff review, and community feedback. At a glance, we observe that users within each country consume content from a mix of international media and local media. Domains from media such as New York Times, Huff Post, CNN, and Reuters are ranked among the top 15 for multiple countries (regardless of English-speaking or not).

From these distributions, we observe considerable audience variations for the same media across countries – in terms of both the mean and spread of audience leanings. For example, *nytimes* is considered a lean-Dem media in the US with a fairly broad audience base in the US, but its audience distribution is much narrower in Germany, and even more so in Australia. In Australia,

lean-Dem media tend to have a narrower audience distribution than center or lean-Rep media. There are also country-specific rankings for media that are distinct from elsewhere. For instance, *thehill* has the second-most lean-Dem audience in Germany, but has a *Center* label and ranks in the middle (by leaning) of the top 15 media for US, Australia and Philippines. *cnn* is considered *L* by MBFC, but its average audience leaning is less to the left than other venues rated *CL*, such as *washingtonpost* and *theguardian* in both the US and Australia.

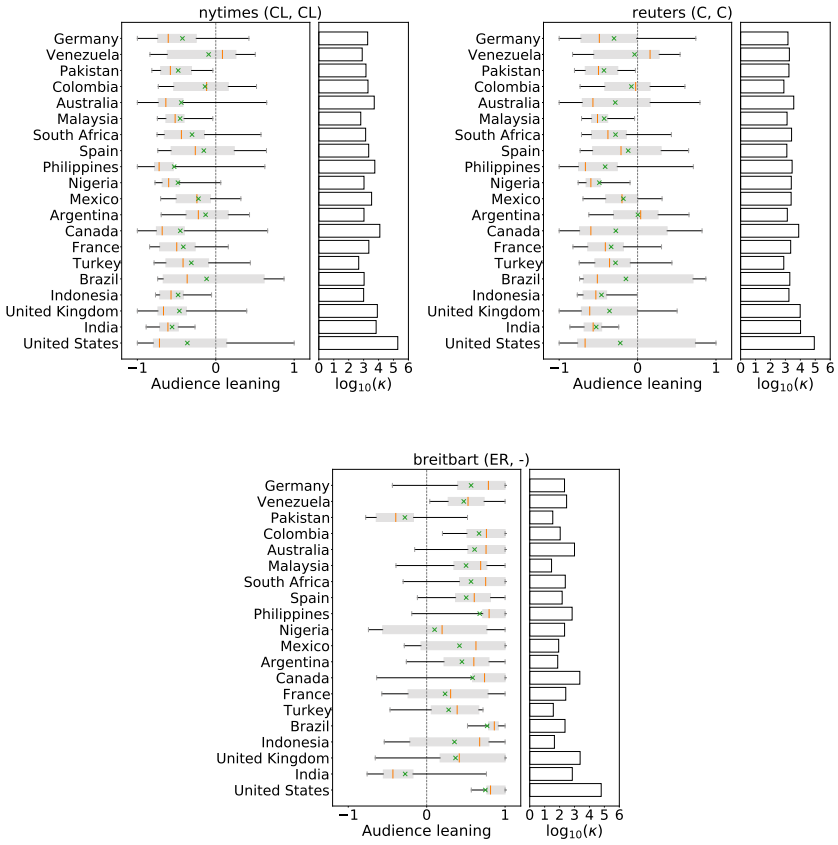


Figure 19: Audience overview across countries for *nytimes*, *reuters* and *breitbart*. For each domain, left subfigure: box plots on audience leaning distribution in 20 different countries. x-axis: audience leaning; y-axis: countries. Green cross: mean; orange vertical line: median. Whiskers span the 10^{th} percentile to the 90^{th} percentile. Right subfigure: bar plots on audience reach in each country. x-axis: audience size in thousands. The MBFC and AllSides labels on the domains are in brackets after each domain name.

A media-centric view. Figure 19 profiles the distribution of audience leaning scores for three example media across the 20 countries. In particular, we choose *nytimes*, *reuters* and *breitbart*, which have been labeled as center-left, center, and extremely-right by MBFC (and Allsides for the first two domains). Figures for more media are in **??**. We see for each media, both the average audience leaning score and the variance of the country-specific leaning scores are very diverse. For instance, *nytimes* has shown to have an average audience leaning score varying between 0 to -0.5,

NPI stance	Precision	Recall	F1 Score	#Identified users
pro-NPI	0.98	1.00	0.99	n=4407863
anti-NPI	0.58	0.27	0.36	n=96177

Table 7: Label propagation performance in 10-fold cross validation for NPI attitude.

and its inter-quartile ranges (width of the box in boxplots) are 0.31 for Australia, 0.94 for the US, but 1.29 for Brazil. For *reuters*, the mean audience leaning is close to 0 in Venezuela, Colombia and Argentina, but 90% of its audience in Malaysia, India and Nigeria are considered lean-Dem. For *breitbart* has an average audience leaning score varying between 0 to 0.75 in most countries. Its audience in US and Brazil are over 90% lean-Rep, whereas in Indonesia, France and Mexico over 25% of its audience are lean-Dem.

These observations are significant due to three reasons. First, while there are many editor-curated [2, 68] and data-driven [10, 21, 78, 44] media bias estimates, all of them focus on a notion of *average* leaning, whereas in this work we quantify the diversity and spread of the audience base. Second, most current leaning results are US-centric [10, 21, 78], or focus on a small number of Western countries [44], our rating is able to cover countries in Asia, South America, Africa, and the Global South at large. Lastly, our data curation and measurement methodology reveal significant cross-country variation in media outlets, the average audience leaning, as well as the political diversity in the audience base.

3.5 Political leaning + Geolocation + NPI

In this section, we move our attention from URL sharing to users' attitudes towards NPI regimes. The attitude towards NPI regime reflects one's adherence/disobedience to various guidelines introduced by governments and health organizations. We are interested in finding out the relations between such adherence/disobedience with one's belief towards the two political parties. We begin this section by reviewing methods for estimating users' attitudes towards NPI regimes. We then present the overview of users along political dimension and NPI dimension. In the end, we break down users' political belief by their attitudes towards NPI.

3.5.1 Estimating users' attitudes towards NPI regimes

We follow a similar approach to political leaning to infer each user's attitude towards NPI regimes within each country. The reduced retweet network has n=4504112 nodes and n=26149407 edges. The performance of label propagation is shown in Table 7.

3.5.2 Global overview of users NPI attitude and political belief

We first assign each user a label for political belief and NPI attitude using the same formula presented in Section 3.2. We then compute the ratio of pro-Democratic/Republican users and pro-/anti-NPI users within each country.

Figure 20a displays the ratio of pro-Democratic/Republican users within the 20 selected countries. The x-axis is the set of countries selected in our study. The y-axis shows the ratio of pro-Democratic users over pro-Republican users in each country. It has shown that each country exhibit

skewness towards pro-Democratic of different levels, i.e. there are more pro-Democratic users than pro-Republican users. Countries like India, Indonesia, Nigeria, Malaysia and Pakistan demonstrates the strongest skewness (ratio > 10). Meanwhile, several countries demonstrate moderate level of skewness such as the U.S., the U.K., Turkey, France, Canada, Philippines, South Africa, Australia and Germany (ratio > 3). There are also countries with mild pro-Democratic skewness such as Brazil, Argentina, Spain, Colombia and Venezuela (ratio < 3).

Figure 20b displays the ratio of pro-/anti-NPI users within the same group of countries. The x-axis is the as Figure 20a. The y-axis shows the ratio of pro-NPI users over anti-NPI users in each country. We can also observe the skewness towards pro-NPI leaning, i.e. a large number of users within each country tend to adhere to the NPI regimes. It is not surprising due to the imbalanced hashtags mentioned in Section 3.2. The figure has shown that countries like India, Indonesia, Turkey, Nigeria, Malaysia and Pakistan have their majority of users in favor of NPI regimes compared to the rest. Meanwhile, countries like U.S., Philippines and Colombia, have more users against such regimes with respect to the whole population in our study.

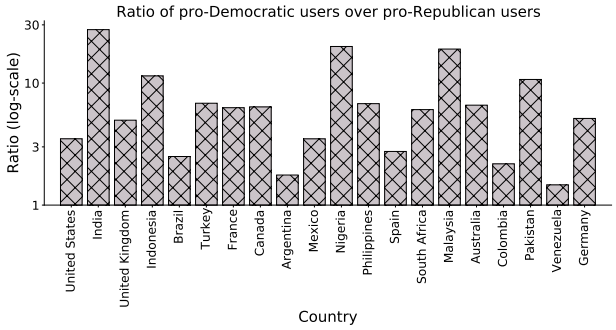
The figures presented here serve as important overview for us to understand the general features of users within each country along the two dimensions. Nevertheless, each country is more than what a ratio could describe. We then present a more detailed analysis over the two topics and the relation between them in the following subsubsections.

From NPI attitude towards political belief In this subsubsection, we present the political leaning breakdown of users based on their attitudes towards NPI regimes. We first partition users within each country into pro-NPI and anti-NPI users. We then compute the ratio of pro-Democratic and pro-Republican users within each group.

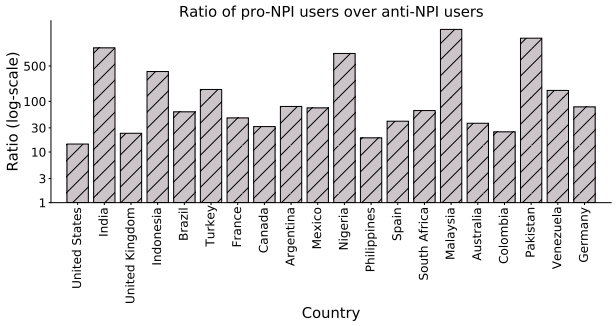
Figure 20c shows the ratio of pro-Democratic over pro-Republican users (Dem-Rep ratio for simplicity) within broken down by their NPI attitude for each country. The x-axis shows the set of countries selected for study. The y-axis shows the two ratio for pro-NPI users (green bars) and anti-NPI users (yellow bars) respectively. Bars growing up implies there are more pro-Democratic users within a particular country.

It is clear that we can observe users' political belief varies across their attitudes towards NPI in most countries. In more than half of the selected countries, Dem-Rep ratio is higher than 1 within pro-NPI users but less than 1 within anti-NPI users. This has implied that within these countries, pro-Democratic users are more likely to adhere to NPI guidelines while pro-Republican users are more likely to disobey the guidelines. Meanwhile, Dem-Rep ratio within anti-NPI users in South Africa and Australia are more extreme than any other countries, indicating a much stronger correlation between political partisanship and attitudes towards NPI guidelines. Recall that the political leaning can also be interpreted as users' viewpoints towards Trump. High Dem-Rep ratio within pro-NPI user groups implies users who follow NPI guidelines tend to oppose Trump. Similarly, low Dem-Rep ratio within anti-NPI user groups implies users who do not follow NPI guidelines are more likely to be Trump supporters.

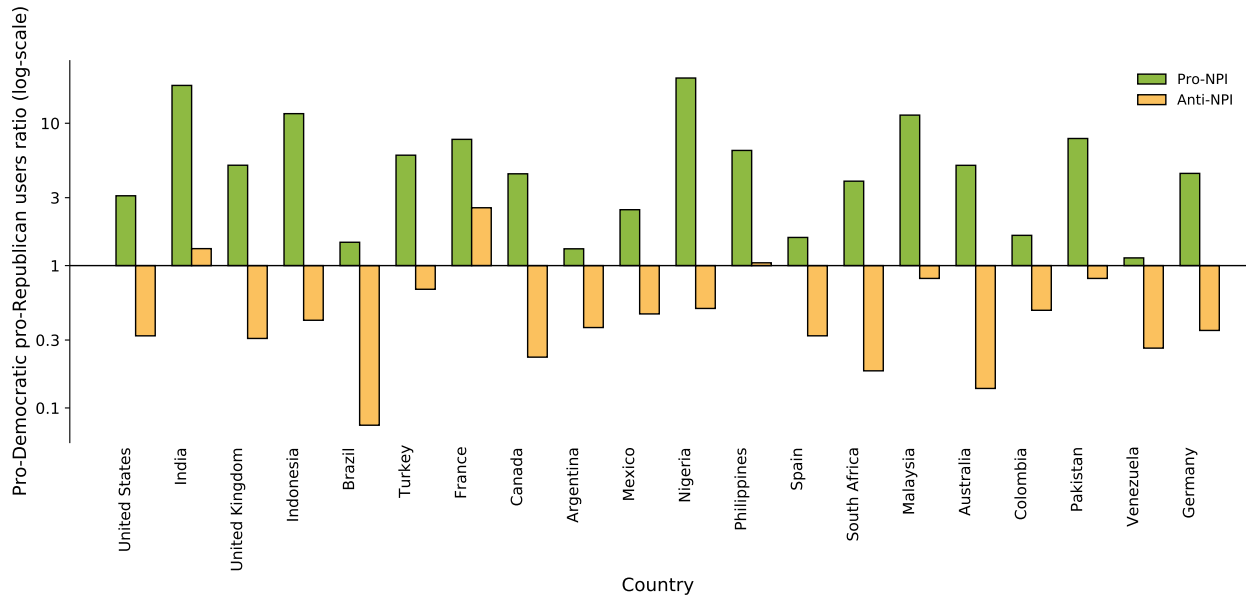
Dem-Rep ratio for both pro-NPI and anti-NPI user groups is higher than 1 in India, France, Philippines, and Pakistan. We attribute it to the pro-Democratic skewness observed previously. In fact, we can indeed see that Dem-Rep ratio for anti-NPI users is still much lower than that for pro-NPI users. It is also surprising to see Argentina, despite suffers less from the left skewness, has more pro-Democratic users than pro-Republican users within anti-NPI user group.



(a) Overall ratio of left-right users within each country. X-axis: countries selected for study; y-axis: ratio for left-right users in log scale.



(b) Overall ratio of pro-anti NPI users within each country. X-axis: countries selected for study; y-axis: ratio for pro-/anti-NPI users in log scale.



(c) Left-right users ratio within pro-NPI users and anti-NPI users. X-axis: countries selected for study; y-axis: ratio for left-right users in log scale. Green bars: left-right ratio within pro-NPI users. Yellow bars: left-right ratio within anti-NPI users. Horizontal line at $y = 1$ represents ratio of 1.

(d) main plot in Sec 7

References

[1] Gagan Aggarwal, Ashish Goel, and Rajeev Motwani. Truthful auctions for pricing search keywords. In *Proceedings of the 7th ACM Conference on Electronic Commerce*, pages 1–7, 2006.

[2] AllSides. Allsides media bias, 2022. https://www.allsides.com/media-bias.

[3] Jisun An, Daniele Quercia, and Jon Crowcroft. Partisan sharing: Facebook evidence and societal consequences. In *Proceedings of the ACM Conference on Online Social Networks*, 2014.

[4] Anonymous. Online supplement for this submission, 2022. https://www.dropbox.com/s/pva532tk3f17gt5/supp_3065.pdf?dl=0.

[5] Sinan Aral. *The Hype Machine: How Social Media Disrupts Our Elections, Our Economy, and Our Health–and How We Must Adapt*. Penguin Random House, 2021.

[6] J. Arrow. An extension of the basic theorems of classical welfare economics. In J. Neyman, editor, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 507–532. University of California Press, 1951.

[7] K. Arrow, H. Block, and L. Hurwicz. On the stability of the competitive equilibrium: II. *Econometrica*, 27(1):82–109, 1959.

[8] Adam Badawy, Emilio Ferrara, and Kristina Lerman. Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2018.

[9] Joseph B. Bak-Coleman, Mark Alfano, Wolfram Barfuss, Carl T. Bergstrom, Miguel Ángel Centeno, Iain D. Couzin, Jonathan F. Donges, Mirta Galesic, Andrew S. Gersick, Jennifer Jacquet, Albert B. Kao, Rachel E. Moran, Pawel Romanczuk, Daniel I. Rubenstein, Kaia J. Tombak, Jay Joseph Van Bavel, and Elke U. Weber. Stewardship of global collective behavior. *Proceedings of the National Academy of Sciences of the United States of America*, 118, 2021.

[10] Eytan Bakshy, Solomon Messing, and Lada A. Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 2015.

[11] Pablo Barberá. How social media reduces mass political polarization. evidence from germany, spain, and the us. *Job Market Paper, New York University*, 2014.

[12] Robert M Bell and Yehuda Koren. Lessons from the netflix prize challenge. *ACM SIGKDD Explorations Newsletter*, 9(2):75–79, 2007.

[13] Michel Benaïm. Dynamics of stochastic approximation algorithms. In *Seminaire de probabilites XXXIII*, pages 1–68. Springer, 1999.

[14] Dimitris Bertsimas, Vivek F Farias, and Nikolaos Trichakis. The price of fairness. *Operations research*, 59(1):17–31, 2011.

[15] Sushil Bikhchandani and John W. Mamer. Competitive Equilibrium in an Exchange Economy with Indivisibilities. *Journal of Economic Theory*, 74(2):385–413, June 1997.

[16] Benjamin Birnbaum, Nikhil R Devanur, and Lin Xiao. Distributed algorithms via gradient descent for fisher markets. In *Proceedings of the 12th ACM conference on Electronic commerce*, pages 127–136, 2011.

[17] Robert M. Bond, Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle, and James H. Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295–298, 2012.

[18] S. Brânzei, R. Mehta, and N. Nisan. Universal Growth in Production Economies. In *NeurIPS 2018*, volume 31, pages 1973–1973, 2018.

[19] Simina Brânzei, Nikhil R. Devanur, and Yuval Rabani. Proportional dynamics in exchange economies. In Péter Biró, Shuchi Chawla, and Federico Echenique, editors, *EC '21: The 22nd ACM Conference on Economics and Computation, Budapest, Hungary, July 18-23, 2021*, pages 180–201. ACM, 2021.

[20] Ceren Budak. What happened? the spread of fake news publisher content during the 2016 us presidential election. In *Proceedings of the Web Conference*, 2019.

[21] Ceren Budak, Sharad Goel, and Justin M. Rao. Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, 2016.

[22] Arthur Capozzi, Gianmarco De Francisci Morales, Yelena Mejova, Corrado Monti, André Panisson, and Daniela Paolotti. Clandestino or rifugiato? anti-immigration facebook ad targeting in italy. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2021.

[23] Ioannis Caragiannis, David Kurokawa, Hervé Moulin, Ariel D. Procaccia, Nisarg Shah, and Junxing Wang. The unreasonable fairness of maximum nash welfare. *ACM Trans. Economics and Comput.*, 7(3):12:1–12:32, 2019.

[24] Emily Chen, Kristina Lerman, and Emilio Ferrara. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance*, 2020.

[25] G. Chen and M. Teboulle. Convergence Analysis of a Proximal-Like Minimization Algorithm Using Bregman Functions. *SIAM J. Optim.*, 3(3):538–543, 1993.

[26] Yun Kuen Cheung, Richard Cole, and Nikhil R. Devanur. Tatonnement beyond gross substitutes? Gradient descent to the rescue. *Games and Economic Behavior*, 123:295–326, 2020.

[27] Yun Kuen Cheung, Richard Cole, and Ashish Rastogi. Tatonnement in ongoing markets of complementary goods. In *EC'12*, pages 337–354, 2012.

[28] Yun Kuen Cheung, Richard Cole, and Yixin Tao. Dynamics of distributed updating in fisher markets. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 351–368, 2018.

[29] Yun Kuen Cheung, Stefanos Leonardos, and Georgios Piliouras. Learning in markets: Greed leads to chaos but following the price is right. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 111–117. ijcai.org, 2021.

[30] Yun Kuen Cheung, Stefanos Leonardos, Shyam Sridhar, and Georgios Piliouras. Market equilibria and risk diversification in blockchain mining economies. In *The 3rd International Conference on Mathematical Research for Blockchain Economy (MARBLE)*, 2022.

[31] Bruno Codenotti, Benton McCune, and Kasturi R. Varadarajan. Market equilibrium via the excess demand function. In *STOC*, pages 74–83, 2005.

[32] Richard Cole and Lisa Fleischer. Fast-converging tatonnement algorithms for one-time and ongoing market problems. In *STOC*, pages 315–324, 2008.

[33] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

[34] Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the ACM Conference on Recommender Systems*, 2016.

[35] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. An experimental comparison of click position-bias models. In *Proceedings of the 2008 international conference on web search and data mining*, pages 87–94, 2008.

[36] Thomas H Davenport and John C Beck. The attention economy. *Ubiquity*, 2001(May):1–es, 2001.

[37] Gerard Debreu. *Theory of value : an axiomatic analysis of economic equilibrium*. Yale University Press, 1972.

[38] Boguslawa Dobek-Ostrowska, Michał Głowacki, Karol Jakubowicz, and Miklos Sükösd. *Comparative Media Systems: European and Global Perspectives*. Central European University Press, 2010.

[39] E. Eisenberg. Aggregation of utility functions. *Management Sciences*, 7(4):337–350, 1961.

[40] E. Eisenberg and D. Gale. Consensus of Subjective Probabilities: The Pari-Mutuel Method. *Ann. Math. Statist.*, 30(1):165–168, 1959.

[41] David S Evans. The economics of attention markets. *Available at SSRN 3044858*, 2020.

[42] Josef Falkinger. Attention economies. *Journal of Economic Theory*, 133(1):266–294, 2007.

[43] Josef Falkinger. Limited attention as a scarce resource in information-rich economies. *The Economic Journal*, 118(532):1596–1620, 2008.

[44] Richard Fletcher, Alessio Cornia, and Rasmus Kleis Nielsen. How polarized are online and offline news audiences? a comparative analysis of twelve countries. *The International Journal of Press/Politics*, 2020.

[45] Santo Fortunato, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Stasa Milojevic, Alexander Michael Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, Alessandro Vespignani, Ludo Waltman, Dashun Wang, and A L Barabasi. The science of science. *Nature*, 138:237, 2018.

[46] Yuan Gao and Christian Kroer. First-order methods for large-scale market equilibrium computation. In *NeurIPS 2020*, 2020.

[47] Daniel C Hallin and Paolo Mancini. *Comparing Media Systems: Three Models of Media and Politics*. Cambridge university press, 2004.

[48] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in neural information processing systems*, pages 1024–1034, 2017.

[49] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

[50] Ferenc Huszár, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. Algorithmic amplification of politics on twitter. *Proceedings of the National Academy of Sciences*, 2022.

[51] K. Jain and V. V. Vazirani. Eisenberg–Gale markets: Algorithms and game-theoretic properties. *Games and Economic Behavior*, 70(1):84–106, 2010. Special Issue In Honor of Ehud Kalai.

[52] Julie Jiang, Emily Chen, Kristina Lerman, and Emilio Ferrara. Political polarization drives online conversations about covid-19 in the united states. *Human Behavior and Emerging Technologies*, 2020.

[53] Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli. Degenerate feedback loops in recommender systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 383–390, 2019.

[54] Kenneth Joseph, Peter M Landwehr, and Kathleen M Carley. Two 1%s don't make a whole: Comparing simultaneous samples from twitter's streaming api. In *International conference on social computing, behavioral-cultural modeling, and prediction*, 2014.

[55] Mamoru Kaneko and Kenjiro Nakamura. The nash social welfare function. *Econometrica: Journal of the Econometric Society*, pages 423–435, 1979.

[56] Jr Kelso, Alexander S and Vincent P Crawford. Job Matching, Coalition Formation, and Gross Substitutes. *Econometrica*, 50(6):1483–1504, November 1982.

[57] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[58] Coco Krumme, Manuel Cebrián, Galen Pickard, and Sandy Pentland. Quantifying social influence in an online cultural market. *PLoS ONE*, 7, 2012.

[59] Haewoon Kwak, Jisun An, Joni O. Salminen, Soon-Gyo Jung, and Bernard Jim Jansen. What we read, what we search: Media attention and public attention among 193 countries. *Proceedings of the 2018 World Wide Web Conference*, 2018.

[60] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the International Conference on World Wide Web*, 2010.

[61] JooYoung Lee, Siqi Wu, Ali Mert Ertugrul, Yu-Ru Lin, and Lexing Xie. Whose advantage? measuring attention dynamics across youtube and twitter on controversial topics. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2022.

[62] D. Levin, K. LaCurts, N. Spring, and B. Bhattacharjee. Bittorrent is an Auction: Analyzing and Improving Bittorrent's Incentives. *SIGCOMM Comput. Commun. Rev.*, 38(4):243–254, 2008.

[63] Haiko Lietz, Claudia Wagner, Arnim Bleier, and Markus Strohmaier. When politicians talk: Assessing online conversational practices of political parties on twitter. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[64] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[65] Spyros Makridakis and Michele Hibon. The m3-competition: results, conclusions and implications. *International Journal of Forecasting*, 16(4):451–476, 2000.

[66] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The m4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, 34(4):802–808, 2018.

[67] Felipe Maldonado, Pascal Van Hentenryck, Gerardo Berbeglia, and Franco Berbeglia. Popularity signals in trial-offer markets with social influence and position bias. *European Journal of Operational Research*, 266(2):775–793, 2018.

[68] MBFC. Media bias/fact check, 2022. https://mediabiasfactcheck.com/mbfc-fact-checks.

[69] Stuart E Middleton, Giorgos Kordopatis-Zilos, Symeon Papadopoulos, and Yiannis Kompatsiaris. Location extraction from social media: Geoparsing, location disambiguation, and geotagging. *ACM Transactions on Information Systems*, 2018.

[70] Amy Mitchell, Jeffrey Gottfried, Jocelyn Kiley, and Katerina Eva Matsa. Political polarization & media habits. *Pew Research Center*, 2014.

[71] Amy Mitchell, Jeffrey Gottfried, Jocelyn Kiley, and Katerina Eva Matsa. Where news audiences fit on the political spectrum. *Pew Research Center*, 2014.

[72] Amy Mitchell, Baxter Oliphant, and Elisa Shearer. About seven-in-ten u.s. adults say they need to take breaks from covid-19 news. *Pew Research Center*, 2020.

[73] Noam Nisan and Ilya Segal. The communication requirements of efficient allocations and supporting prices. *Journal of Economic Theory*, 129(1):192 – 224, 2006.

[74] Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-beats: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*, 2020.

[75] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*, 2017.

[76] refinery29.com. Two political parties have put decriminalising abortion in their manifesto. https://www.refinery29.com/en-gb/2019/11/8876802/abortion-uk-manifesto-election-2019, 2019. [Online; accessed Jul-15-2022].

[77] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[78] Ronald E Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. Auditing partisan audience bias within google search. *Proceedings of the ACM on Human-Computer Interaction*, 0(CSCW), 2018.

[79] Matthew J. Salganik, Peter Sheridan Dodds, and Duncan J. Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311:854 – 856, 2006.

[80] M. Ángeles Serrano, Marián Boguñá, and Alessandro Vespignani. Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences*, 2009.

[81] Sono Shah, Emma Remy, and Aaron Smith. Differences in how democrats and republicans behave on twitter. *Pew Research Center*, 2020.

[82] Minjeong Shin, Alasdair Tran, Siqi Wu, Alexander Mathews, Rong Wang, Georgiana Lyall, and Lexing Xie. Attentionflow: Visualising influence in networks of time series. In *The 14th International Conference on Web Search and Data Mining, Demo*, WSDM '21, 2021.

[83] Herbert A Simon et al. Designing organizations for an information-rich world. *Computers, communications, and the public interest*, 1971.

[84] Hiroki Takikawa and Kikuko Nagayoshi. Political polarization in social media: Analysis of the "twitter political field" in japan. In *2017 IEEE international conference on big data (big data)*, pages 3143–3150. IEEE, 2017.

[85] Lester C Tong, M Yavuz Acikalin, Alexander Genevsky, Baba Shiv, and Brian Knutson. Brain activity forecasts video engagement in an internet attention market. *Proceedings of the National Academy of Sciences*, 117(12):6936–6941, 2020.

[86] Alasdair Tran, Alexander Mathews, Cheng Soon Ong, and Lexing Xie. Radflow: A recurrent, aggregated, and decomposable model for networks of time series — supplementary materials. https://github.com/alasdairtran/radflow, 2021.

[87] Alasdair Tran, Alexander Patrick Mathews, Cheng Soon Ong, and Lexing Xie. Radflow: A recurrent, aggregated, and decomposable model for networks of time series. In Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia, editors, *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 730–742. ACM / IW3C2, 2021.

[88] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *ArXiv*, abs/1706.03762, 2017.

[89] Petar Velivcković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

[90] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[91] L. Walras. *Éléments d'économie politique pure ou théorie de la richesse sociale (Elements of Pure Economics, or the theory of social wealth)*. Routledge, Lausanne, Paris, 1874. (1899, 4th ed.; 1926, rev ed., 1954, Engl. transl.).

[92] F. Wu and L. Zhang. Proportional Response Dynamics Leads to Market Equilibrium. In *STOC '07*, pages 354–363, 2007.

[93] Neo Z. Wu, Bradley A. Green, Xue Ben, and Shawn O'Banion. Deep transformer models for time series forecasting: The influenza prevalence case. *ArXiv*, abs/2001.08317, 2020.

[94] Siqi Wu and Paul Resnick. Cross-partisan discussions on youtube: Conservatives talk to liberals but liberals don't talk to conservatives. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2021.

[95] Siqi Wu, Marian-Andrei Rizoiu, and Lexing Xie. Beyond views: Measuring and predicting engagement in online videos. In *Twelfth international AAAI conference on web and social media*, 2018.

[96] Siqi Wu, Marian-Andrei Rizoiu, and Lexing Xie. Estimating attention flow in online video networks. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW):183:1–183:25, November 2019.

[97] Siqi Wu, Marian-Andrei Rizoiu, and Lexing Xie. Variation across scales: Measurement fidelity under twitter data sampling. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2020.

[98] Cai Yang. Measuring covid-19 related media consumption on twitter, 2021. Bachelor Thesis, The Australian National University.

[99] Cai Yang, Siqi Wu, and Lexing Xie. The shapes of the fourth estate worldwide: Profiling global media audience during the covid-19 pandemic, 2022. Under revision at ACM CSCW, Oct 2022.

[100] L. Zhang. Proportional response dynamics in the Fisher market. *Theor. Comput. Sci.*, 412(24):2691–2698, 2011.

[101] Ling Zhao, Yujiao Song, Chao Zhang, Yu Liu, Pu Wang, Tao Lin, Min Deng, and Haifeng Li. T-gcn: A temporal graph convolutional network for traffic prediction. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–11, 2019.

[102] Dengyong Zhou, Olivier Bousquet, Thomas Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. *Advances in neural information processing systems*, 2003.

[103] Haiqing Zhu, Yun Kuen Cheung, and Lexing Xie. Stability and efficiency of personalised cultural markets, 2022. Under review at ACM TheWebConf2023.

[104] Kai Zhu, Dylan Walker, and Lev Muchnik. Content growth and attention contagion in information networks: Addressing information poverty on wikipedia. *Information Systems Research*, 2020.