News Reading Publics and their Evolution in Multi-Lingual Political Contexts: Evidence from Online India (2014-2018)

Subhayan Mukerjee[1*], Silvia Majo-Vazquez[2], Sandra Gonzalez-Bailon[1], and Rasmus Kleis Nielsen[2]

[1]Annenberg School for Communication, University of Pennsylvania
[2]Reuters Institute for the Study of Journalism, University of Oxford
[*]*corresponding author: Subhayan Mukerjee (subhayan.mukerjee@asc.upenn.edu)*

Abstract

In this paper, we propose a theoretical framework called news reading publics to understand news consumption as a global socio-political process. we first draw from a variety of existing communication theories and argue that most Western-centric research on news consumption paints a parochial picture by not focusing on the cultural aspect of the consumption process. we next describe the use of a network-based methodology to theorize the existence of these news reading publics in a culturally variegated context, validate it using a novel dataset of online audiences from India and demonstrate its ability to offer a rigorous comparative framework for studying audiences across diverse countries.

Keywords— news consumption; network science; news audience; online news

1 Introduction

News consumption has been one of the mainstays of daily life since the irruption of mass newspapers in the 18th century [27, 47] and it plays a central role in the relationship between

citizens and the political process in any country [10]. While news consumption has, at its core, remained the same over the past several decades, satisfying a fundamental societal need and enabling the political process, it has witnessed several structural changes in its evolution that have transformed the manner in which it fulfills these functional requirements. These structural changes have been the effect of technological innovations on the one hand, and their subsequent reception and interpretation by the mass publics on the other. Thus, the landscape of news consumption has constantly adapted to reflect not just the state of mass media at any point in time, but also the organization of its audiences. The Internet has been the latest in a long line of technological breakthroughs that have had an indelible effect on this landscape.

In parallel with the evolution of media systems, the manner in which new technological affordances impacted the process of news consumption and in turn impacted society, has dominated media scholarship for several decades. Moreover, because these technological innovations originated predominantly in the US, these academic inquiries were also largely rooted in American contexts and colored with American perspectives. Therefore, when towards the end of the 20th century, the rise of cable television and the birth of Fox News and MSNBC raised concerns over audience fragmentation, it inherently implied fragmentation in a manner that related to an American-only audience, in other words, along Democratic-Republican partisan lines [38]. Similarly, the specialization of news audiences has also predominantly been thought of along demographic and generational terms. These lines of thinking have subsequently permeated audience research in the internet age as well. As a result, a plethora of studies since the turn of the 21st century, that seek to understand how people consume news online have made the simplifying connection with partisan identity and political polarization, not just in the United States (for egs. [1, 14, 26, 31, 34], but elsewhere as well [9].

In this paper, I argue that these traditional approaches to understanding audiences severely limit their applicability, especially when used in non-American contexts because they paint an incomplete picture of how news consumption actually operates. I argue that for truly global audience research, there is a need for a holistic conceptual framework that can scale beyond a single country and help inform frameworks that can be used to compare audience dynamics across countries. Next, I use existing social scientific theories in communication and psychology to introduce such a framework. Finally, I use a novel dataset tracking online news consumption in India and network analytic methods to formalize the theoretical intuitions behind this framework. To conclude, I discuss how the same method can be applied elsewhere and be used for methodologically robust comparative research.

2 News Reading Publics as a Networked Framework

One of the reasons why India is a useful foil to typical western democracies from the perspective theory building is its high level of cultural heterogeneity. This makes it a potentially useful candidate for informing an intuitive framework about news consumption that goes beyond the narrow conception of culture that has dominated political communication research in recent years - one that is primarily partisan [12]. However, India is not unique in this regard, and any substantially diverse and variegated country or population could serve as a similarly useful substrate. Yet, India is the largest democracy in the world, and continues to be woefully under-represented in audience research. Therefore, the case of India is "not just intrinsically important, but also of broader relevance as an example of a market in the global South" [28], and theory building informed by empirical evidence from such a context is crucial if the field of communication is to be truly global.

While as a democracy, India shares several political apparatuses and institutions with other democracies in the West (like the U.K. and other Commonwealth countries), its cultural heterogeneity greatly affects how the political process plays out in reality. The regionally variegated and poly-centric media environment [7] plays a central in this regard. In fact, one of the primary differences of the Indian media from traditional media environments in Western countries stems from the extent of the linguistic diversity in the country. As a result, the different states in India, each of which has its own language , also have thriving regional media industries in their respective vernaculars. A very eloquent explanation of the term "regional media" in the context of India is offered by Shanti Kumar [23]: "The term regional media is traditionally used to describe an intranational or subnational category of media produced in the many regional languages spoken within India. Media produced in Hindi – particularly in cinema and television – are considered "national" since Hindi is the "national" language. But in other contexts, Hindi media – particularly in print journalism – are considered "regional" in contrast to English media, which circulate predominantly in urban areas and are extremely influential among the political, economic, and cultural elites in the metropolitan centers. Since most states within India are linguistic states, regional media in India are usually defined in relation to the geographic boundaries of the states and the dominant languages within those states".

#	Language	Native Speakers	% of Total Population
1	Hindi	528347193	43.63
2	Bengali	97237669	8.03
3	Marathi	83026680	6.86
4	Telugu	81127740	6.7
5	Tamil	69026881	5.7
6	Gujarati	55492554	4.58
7	Urdu	50772631	4.19
8	Kannada	43706512	3.61
9	Odia	37521324	3.1
10	Malayalam	34838819	2.88
11	Punjabi	33124726	2.74
12	Assamese	15311351	1.26
13	Maithili	13583464	1.12
14	Santali	7368192	0.61
15	Kashmiri	6797587	0.56
16	Nepali	2926168	0.24
17	Sindhi	2772264	0.23
18	Dogri	2596767	0.21
19	Konkani	2256502	0.19
20	Manipuri	1761079	0.15
21	Bodo	1482929	0.12
22	English	259678	0.02
23	Sanskrit	24821	~ 0

Table 1. Official Languages as recognized by the Indian Constitution and the number of native speakers. Source: Census 2011

One can now begin to have certain expectations about how the existence of these distinct regional media systems in parallel with the normatively understood "national media" will be reflected in the news use patterns at the individual level. This is because Indian citizens tend to identify with their state as well as their country, both culturally and politically. An example of the difference in cultural identities lies in the celebration of festivals. Every state has its own set of holidays corresponding to their respective festivals and cultural events, in addition to a nationally mandated set of holidays like the Independence Day and Republic Day. An example of the difference in political identities can be found in political coalitions

and alliances formed by regional political parties with national parties (and other regional parties). It is, therefore, not unnatural for an Indian citizen to vote for a certain party at the state assembly elections, but for an ideologically very different party at the general (national) elections.

These cleavages that arise between the local – "the state" and the national – "the country" in India are echoed in the power-dynamics between the regional media and the national media as well – with either media having to deal with overlapping but often distinct issues to report on and very different expectations from consumers to handle. This points to the existence of different kinds of publics in the India citizenry – comprising people who share certain identities and issues with some others, but not with everyone. A person living in Mumbai, Maharashtra for instance would probably not care too much about the issues facing those who live in the state of West Bengal (for example, the Gorkhaland separatist movement, or immigration across the Bangladesh-India border), simply because they are of very little relevance to them. Thus, outside of certain issues that are of national importance (for example, the national elections, foreign policy or the national budget), the regional media industry in Maharashtra would need to focus on a very different set of media requirements for the people of its state than the regional media industry in West Bengal (for example, sharing of river water with neighboring states, or state elections). The divide is further exacerbated by the presence of different regional political parties and the difference in language between the two states. The average news consumer in India, in other words, can be thought of as playing two "roles" – the first being that of a consumer of "regional" news, and the second, as a consumer of "national" news. The empirical research question that then arises is this: to what extent is this duality of consumer patterns reflected in how Indians navigate news sources on the internet? In other words, how can we identify these different publics by studying news consumption patterns?

It is by examining the data and answering this question that I propose a networked theory of news reading publics – a theoretical framework to help understand the organization of multiple kinds of online audiences, each with a unique set of expectations and consumption patterns yet sharing heavy commonalities with all others. In the next section, I describe how such a framework can be thought of to be the natural extension of several existing theoretical frameworks by applying them to the context of news consumption. I consolidate these multiple existing lines of theoretical inquiry – that of uses and gratifications, social identity theory, the cultural logic of consumption, and issue publics – and help pave the way towards a general framework for understanding media audiences. Moreover, despite it being empirically grounded in the context of a pluralistic and non-Western democracy like India, such a framework can also offer a theoretical lens to compare news consumption

patterns across countries, including those in the west, in the future.

3 Situating News Reading Publics in Communication Theory

3.1 Uses and Gratifications

Communication as a discipline, witnessed a paradigmatic shift in the mid-1950s, in rethinking about how audiences engaged with media. This new line of thinking gave individuals some agency in how they dealt with messages. Attribution theory [18, 22, 49] for instance, thought of understanding media effects by attributing individuals with feelings, beliefs, and intentions. Scholars theorized that different preexisting cognitive orientations of different individuals could lead to different levels of acceptance of media messages. Thus, two individuals were no longer equal as far as media effects, and more generally media consumption, was concerned, and the manner in which they engaged with the information, depended on their individual selves, not simply on the message. Individuals were thus part of an "active" audience [3].

The Uses and Gratifications (UG) model was one of the logical descendants of the active audience theory. This framework went one step beyond attribution theory and hypothesized that individuals were not just receiving the same messages in cognitively different ways, but they were actively seeking out (or blocking out) the media messages that they wanted Blumler1974, Katz1973. As opposed to asking what the "media does to people", a UG approach focused on asking "what do people do with media" [21].

In the context of news consumption, the UG approach helps reconceptualize audience behavior in a more structurationist manner [15]. It lends itself particularly well to the variegated media environment in a country like India, because the linguistic differences in the different parts of the country create a nested media structure that is significantly determined by the linguistically different media demand that exists in the various regions of the country. It is for the same reason that English newspapers are circulated throughout the length and breadth of India, but vernacular newspapers are not. The very first Hindi newspaper, for instance, was the Udant Martand, and it began publication in 1826 from Calcutta, the then capital of British India. However, Calcutta, in West Bengal, being a Bengali speaking city in a Bengali-majority state, had a very small market for news in Hindi. Thus, the newspaper struggled to sustain itself and finally stopped publication the very next year.

In other words, if we use language as a marker, the Indian media consumption landscape can be thought of as one that is shaped to a great extent by a linguistic familiarity gratification. Moreover, because language encodes cultural and regional differences that run much deeper, one can expect this to be symptomatic of a more general gratification of familiarity: i.e. people seek out and consume news they are familiar with and block out news they are not. We can now hypothesize that Indian citizens consume national news and regional news because both these consumption activities satisfy this gratification – a familiarity with the regional, and a familiarity with the national. Thinking about Indian news audiences using a UG framework thus helps understand why regional media have thrived in the various parts of the country alongside English media that has operated at a national scale. It helps us to begin conceptualizing the duality of identity of Indian citizens as news consumers that I had briefly mentioned in an earlier section. This gratification of familiarity (which translates into two parallel demands for regional media and national media in the country) thus reflects a complex dynamic of identities and social identities that characterizes an individual Indian citizen.

3.2 Identity and Social Identity Theory

Identity theory [30, 42, 43, 44] attempts to explain social behavior in terms of the relationship between the self and society. It "views the self not as an autonomous psychological entity but as a multifaceted social construct that emerges from people's roles in society" [19]. According to Stryker, every individual has distinct components of self, each of which is a role identity, that corresponds to each of the role positions they occupy in a society (Ibid.). Two concepts that are closely related to this theory are identity salience and commitment. The former refers to the importance that an individual assigns to each of their identities, thereby creating a hierarchy of identities within one's self [29], while the latter refers to the manner in which they choose the different levels of importance, thereby committing to varying degrees, to each of their identities [45].

Social identity theory, on the other hand, focuses more on understanding inter-group relations. This theory stems from the idea that "a social category (e.g., nationality, political affiliation, sports team) into which one falls, and to which one feels one belongs, provides a definition of who one is in terms of the defining characteristics of the category" [19]. To explain how this operates, social identity theorists invoke the mechanisms of self-categorization and self-enhancement. The former helps create more stereotypical prototypes of groups as well as produces normative perceptions and actions, which in turn assigns individuals to a specific social category (see also self-categorization theory: [48]. The latter focuses on the

need of the individual belonging to a social category, to see and present themselves in better light to other in-group members.

In the context of news consumption, individuals indicate their different levels of salience for their different identities by actively seeking news and information about issues and topics that they deem relevant in their hierarchical conception of their selves. News consumption can thus be understood to be a behavioral manifestation of their role identities. The higher one positions their identity in a particular role in society within their self, the more likely they are to seek information pertaining to that role.

Similarly, from the perspective of self-categorization theory, the consumption of in-group relevant information constitutes one of the characteristic features of a group. This behavior, however, also potentially has a self-enhancement aspect to it arising out of social desirability: people tend to consume more news that is relevant to their in-group in order to appear knowledgeable about the group to other in-group members.

For many Indian citizens, who also strongly identify with their respective states, it is easy to see why they would take on two roles as news consumers: the first being a consumer of regional news, and the second, being a consumer of national news. This existence of multiple news consuming roles however, is not specific to India, even though it may be more prominent in India than in a country like the US. More generally, every individual in any society plays multiple news consuming roles (each corresponding to an identity): what makes the case of India useful is that these roles can be empirically identified, isolated, and analyzed from large scale news consumption patterns. This can crucially help in complementing the theory with a set of methods and measures – which I introduce later in this paper.

3.3 Issue Publics

The concept of issue publics provides yet another lens to understand how news audiences self-organize. First coined by Philip Converse Converse1964, the term has come to mean groups of citizens who "who follow particular issues with close and relatively continuous attention" [37]. Moreover, like any "public" defined in the sociological sense, it does not refer to any specific geopolitical entity; instead it is used to define a loosely organized collective, formed out of a shared interest in specific issues [36]. The concept of issue publics is particularly useful from the perspective of news consumption research because it helps understand commonalities within different parts of a news consuming population.

In the context of India, issue publics can be theorized to be split along the political divisions of the state boundaries, while also simultaneously existing at a national level. This

also maps on to the dual identity of Indian citizens discussed above. Because the political system in India is characterized by a large number of regional parties that only operate within certain states, as well as national parties that operate throughout the country, political platforms that parties run on for elections and political issues that voters (theoretically) vote on, can be very distinct and diverse in different states. Moreover, the cultural differences between states also makes various aspects of daily life very different. For example, different states celebrate different festivals and observe different sets of holidays. These variations in political and cultural issues across the country give rise to regionally fragmented issue publics, who, during the process of news consumption, seek out news that is of specific relevance to them. These different regional issue publics, however, are also nested within a national issue public – because as citizens of India, people living in the various states are also interested in issues relevant to the country at large. Examples of such issues could be the general parliamentary elections, parliamentary proceedings, and foreign policy among others.

3.4 Cultural Proximity

Our final strand of theory for developing the framework of news reading publics comes from the cultural communication literature. In the previous chapter, we discussed how the Indian media environment has historically been shaped by forces of media imperialism, and subsequent Indianizing pushbacks. Cultural theorists argue that the mechanism underpinning these dynamics lies in an indigenous preference for cultural proximity that acts as a natural gatekeeper for foreign media in non-Western countries: "viewers tend to actively privilege national or regional programming over its imported counterpart and in fact rarely turn to imported programs when local alternatives are available" [6]. In other words, when Western imperialistic powers (either nation states or corporates) enter new markets, they encounter local competition as the market saturates. At that point, audiences in these markets tend to automatically prefer to consume the locally produced content to the Western content because of their cultural proximity to the former [8].

This is in some ways an extension of the uses and gratifications framework, which underscores the element of agency that enables audiences to choose what media they want to consume [4]. Other cultural effects studies (ones, for example, done in Saudi Arabia [5], South Korea [20], and the Philippines [46]), have also found how "Western media differentially affected foreign audiences" [41].

In the particular context of India, a preference for cultural proximity is reflected in the division of consumption patterns between English media and vernacular media. Rajagopal

[40] writes: "the historical cleavage between English as the language of command and the indigenous languages was accentuated with independence and the new elite. The usage of English became the shortcut and compensation for the absence of a cultural policy, stringing a high wire above the particularist thickets of region, caste and religion. To summarize a complex argument here, the English language, by virtue of being subsumed as [a] language of command, continued a colonial practice of aloofness and unfamiliarity with local traditions." Thus, for the swathes of Indians living in rural areas English media, despite being "national", is simply too unfamiliar, when compared to the more "culturally proximal" vernaculars. This in turns points to the existence of theoretically distinct news reading publics – defined, yet again, in terms of cultural familiarity and language.

A theory of news reading publics aims to consolidate these various theoretical approaches to try and create a general and universally applicable framework for understanding audiences based on shared commonalities. By analyzing the temporal evolution of these news reading publics, one can better characterize the overall evolution of the news consumption landscape in terms of the relative prominence of these publics. For example, it can help answer questions pertaining to certain news reading publics becoming more or less prominent over time, and what these trends mean for the health of the news media in general.

4 Data and Methods

The empirical contribution in this paper relies on a novel dataset obtained from Comscore that tracks how audiences consume news in India online. Comscore is a media analytics company that tracks online traffic in more than forty countries for market research – including India. The panels they maintain in each of these countries are representative of the respective online populations. The recruited panelists are required to install a piece of software on their computers which then passively tracks their web browsing activity. Comscore also has agreements with major media outlets in different countries that allows them to embed tracking tags in source code of their respective websites. These tags can record the number of times each of these websites is visited and, crucially, because they operate on the server side, they are able to capture total website visits, not just of those individuals who are on the Comscore panel. The trace data that are obtained from the browser activities of the panelists is then integrated with the server-side web traffic records that are obtained from the embedded tags using a proprietary algorithm. Upon integration, Comscore is able to generate reliable monthly estimates capturing various aspects of browsing activity – for instance, the number of unique page views every month, the average time spent on each web-page every month, and so on.

There are two important things to note about the data: the first, is that the media websites that are included in each month in the dataset are those with at least 0.1% reach of the total unique visits captured in that month's data. The second is that the data only reflect desktop use and not mobile use. While the omission of mobile traffic data may be a cause for concern, it is important to note that the phenomenal growth of mobile in India is largely driven by consumption in the categories of 'entertainment' (27.38%) and 'sports' (18.64%) media. All other forms of media consumption on mobile platforms are still quite low. For instance, only 5.12% of media consumption in the 'politics' category, and 1.62% in the 'economics' category happens on mobile [33]. My focus on online news audiences as opposed to the general online audience can, therefore, partially help circumvent this limitation. Finally, it is also imperative to note that Comscore's estimates are the best available dataset for India.

4.1 Variables and Measures

The Comscore dataset has two statistics that are of relevance. The first is audience reach, which gives the monthly estimated number of unique visitors for every web page. The second is cross-visiting, which gives the monthly estimated overlap of unique visitors for every pair of web pages. Thus, while audience reach for website A for the month of December 2017 can tell us how many unique individuals visited website A in December 2017, the cross-visiting estimate for websites A and B can tell us how many unique individuals visited both websites A and B in December 2017. These data are available over a period of 45 months – from October 2014 to June 2018.

Details of the methodology are elucidated in the next section (see Data Preparation), but very broadly, it involves the construction of monthly networks of audience overlap, where the nodes are websites, and the edges between them denoting the strength of the corresponding shared audience. I then add additional node attributes to every network, that capture the nature of the media outlet the node represents. These are categorical variables and are as follows: Regional (ie. whether an outlet is regional or national), Digital-born (ie. whether an outlet is digital-born or has been in existence since before the Internet), Indian (ie. whether an outlet is Indian or foreign), and Language (ie. whether an outlet is published in English, Vernacular, or in both).

The logic behind the addition of these variables is to be able to study the differences in how they are embedded within the networks, and to help answer the questions related to their power-differentials in commanding consumer attention – which in turn can inform the theory of news reading publics that I proposed earlier. Yet other variables – that capture

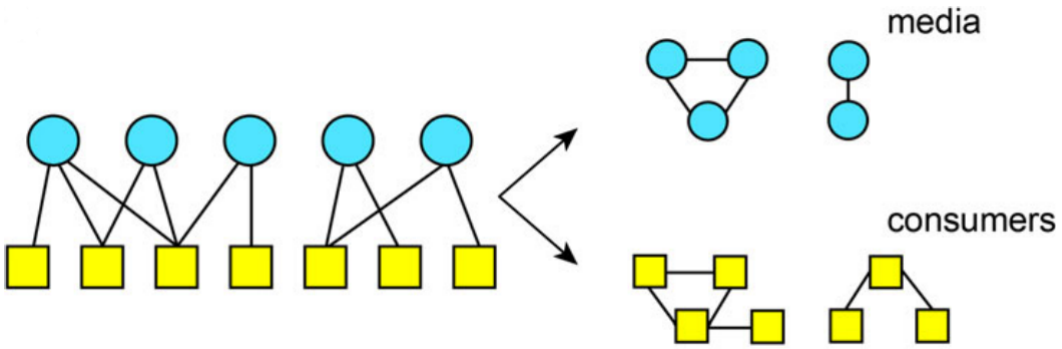

Figure 1: The construction of audience networks (Mukerjee et al., 2018a). The blue circles denote media outlets, and the yellow squares, media consumers. The media outlet-media consumer bipartite network can be split into two projections, one of which is the network of media outlets. This projection is what we construct from the audience overlap data that Comscore gives us access to.

various features of the networks (both at the monthly or individual level, as well as at an aggregated level) like modularity, centralization, and inequality – are detailed in the analysis sections.

4.2 Data Preparation

The first step of my analysis is the creation of networks of media consumption for each of the 45 months using a previously published methodology [32]. To do this, I treat each media outlet as a node that is connected to other media outlets based on the cross-visiting data or the "audience overlap" for each pair. The networks are therefore undirected and weighted, with the weight of edges between nodes capturing the strengths of audience overlap between them. The logic of network construction is detailed in Figure 1.

Once the networks are built, the audience reach of each of the individual media outlets is assigned to the corresponding media node as a node attribute. The total number of media outlets captured in the dataset is 352; however, since every website does not qualify for the 0.1% criteria for percent reach every single month, the number of websites/media outlets that are recorded in the dataset every month tends to vary. Consequently, the network size as measured in terms of the number of nodes in the network also varies from month to month. Thus, for the sake of comparability and uniformity, I only preserve those nodes in the networks that appear in all the 45 months. The final number of nodes that remain the data for each of the 45 months is 177.

Next, I filtered the network so as to only retain the edges that are statistically redundant following prior work [32]. This edge reduction method benefits us in two ways: they significantly reduce the number of edges in the networks and thus make analyzing the networks computationally less intensive; more importantly, they both do a better job of removing noise than other edge reduction methods owing to their use of statistical benchmarks as opposed to an arbitrary threshold parameter.

4.3 Methods

The existence of various thriving regional media industries in India is in sharp contrast to media industries in western democracies like the US, because unlike in the US, local news in India, is neither at risk and nor in decline. In other words, the regional media industries have held their ground, and prevented being subsumed by the big national media houses. As has been mentioned earlier, the success of the regional media industries has largely been driven by the cultural and political divisions between the national and the state levels, which has consistently maintained a market for local news in parallel with a market for national news, in almost every single state. Thus, the Indian news consumer plays the dual role of "regional news consumer", as well as "national news consumer" in the consumption landscape. If this were indeed true, then what evidence of such behavior would we expect in our Comscore dataset? The answer lies in the application of a very specific network intuition – that of *community detection* and its related network metric, *modularity*.

A community in network science refers to a subgraph of a network that is characterized by dense connections between nodes within itself but relatively sparser connections with nodes outside it. This intuition is captured by the network metric modularity – that when split into different communities or modules, the number of edges between nodes within the same module is higher than what is expected by chance when the edges in the same network are randomly wired. Its value lies between -1 and 1, with values above 0 generally signifying that the network is more modular than a randomly rewired network with the same number of nodes and edges. The concept of community detection stems directly from the idea of modularity and network modules. Community detection is the name given to a family of algorithms that attempt to partition a given network into these communities [24, 25]. One class of these algorithms tries to achieve it by iterating through all possible sets of community partitions, and finally converging on one that maximizes the modularity of the network. These community detection algorithms are also known as "modularity optimization" algorithms. One of the most widely used common community detection algorithms is the Walktrap algorithm [35]. This algorithm operates under the intuition that

a person walking randomly along the edges of a network, over a sufficiently long period of time, will tend to get trapped within the communities of the network, were they to exist.

A method for parameterizing community detection algorithms, for more fine-grained detection of modular substructures was demonstrated by Arenas et al. (2008). To do this, the authors added a self-loop (i.e. an edge beginning at and ending on the same node) of a fixed weight to every node of the network, to boost the strength of each node while retaining the network's overall topographical structure. The weight of the self-loop was a tuning parameter, and by varying it across the parametric space, the authors were able to get different modular structures at different resolution levels of the networks. The qualitative intuition behind this method as applied to the Walktrap algorithm is that by increasing the strength of the nodes, the random walker is more likely to remain at a node upon reaching it (and less likely to walk to an adjacent node), thereby restricting her overall mobility and yielding a different outcome. I borrow this intuition with one crucial alteration. Instead of assigning a self-loop with the same weight to every node, and tuning this value over the parameter space, I use the node's audience reach as the weight of its own self-loop. Thus, every node now has a self-loop with a weight unique to itself, that captures the theoretical notion of the node strength, better than a randomly varying parameter would.

The ideas of modularity and community detection can be very useful for testing the duality of news consumption patterns in India. In the audience overlap network, one would therefore expect communities that represent the various vernacular regional media that exist in India. This is because media outlets from the same region of the country would naturally tend to have higher audience overlap than media outlets from different parts of the country. Moreover, these regional communities would also have strong links to the national media nodes. A community detection algorithm that tries to optimize the modularity of the network should therefore be able to find communities in the network that a) represent various distinct regional media in India, and also b) represent the national media. Moreover, there should yet be another community that reflects other potentially specific news reading publics beyond just the national and regional ones: for instance, international news consumers. This is because international news (for example, the BBC or The New York Times) appeals to a very specific demographic of the Indian middle class – those that are young, English educated, and urban – a section, despite being small when compared to the entire Indian citizenry, is a substantial section of the online public in India. I hypothesize the presence of linguistically homogenous (regional) communities, each of which will share heavy overlap with a national media community, but minimal overlap with each other and with an international media community, in the network. Of the linguistic communities, I expect stronger overlap between North Indian language communities (owing to the prepon-

derance of Hindi speaking people in these regions), and greater fragmentation among the South Indian communities owing to the lack of any particular dominating language in these states. These communities can be understood to be distinct yet overlapping news reading publics. I also suspect the international media community to share strong connections with the national media communities because those who are more likely to consume international news are also likely to belong to the national news reading public. The strong overlap between the regional and national communities but minimal overlap between different regional communities would serve to lend credence to the dual-role theory of the Indian news reading public that has been asserted earlier.

Once these news reading publics are identified using the community detection methods, comparing the temporal trends of the different news reading publics will provide insights into whether the news consumption landscape in India is characterized by the increasing dominance of certain publics vis-à-vis others in so far as the 45 month period is concerned.

5 Findings

5.1 Descriptive Statistics

The breakdown for the number of media outlets by type is depicted in Figure 2. These tables tell us that there is a definite skew towards national outlets, towards legacy outlets, and towards English outlets in the online Indian news consumption landscape.

The panels A through I in figure 3 compares the descriptive statistics of the raw networks, the networks after preserving only the outlets that occur every month (induced), and finally, after thresholding using the dyadic null model (filtered).

Panels A through D in figure 4 show scatterplots of the percentage reach and the degree centrality of each of the 177 news outlets based on type, averaged over the 45-month period. The plots are done on log-log axes to reduce the skew in the data for visualization purposes. The Spearman's rank correlation between the percentage reach and the degree centrality for the 177 outlets is 0.908. As can be seen from the visualizations, English language outlets tend be more central, and have greater reach than vernacular outlets, Indian media outlets tend to be more central and have greater reach than foreign media outlets, and national media outlets tend to be more central and have greater reach than regional media outlets.

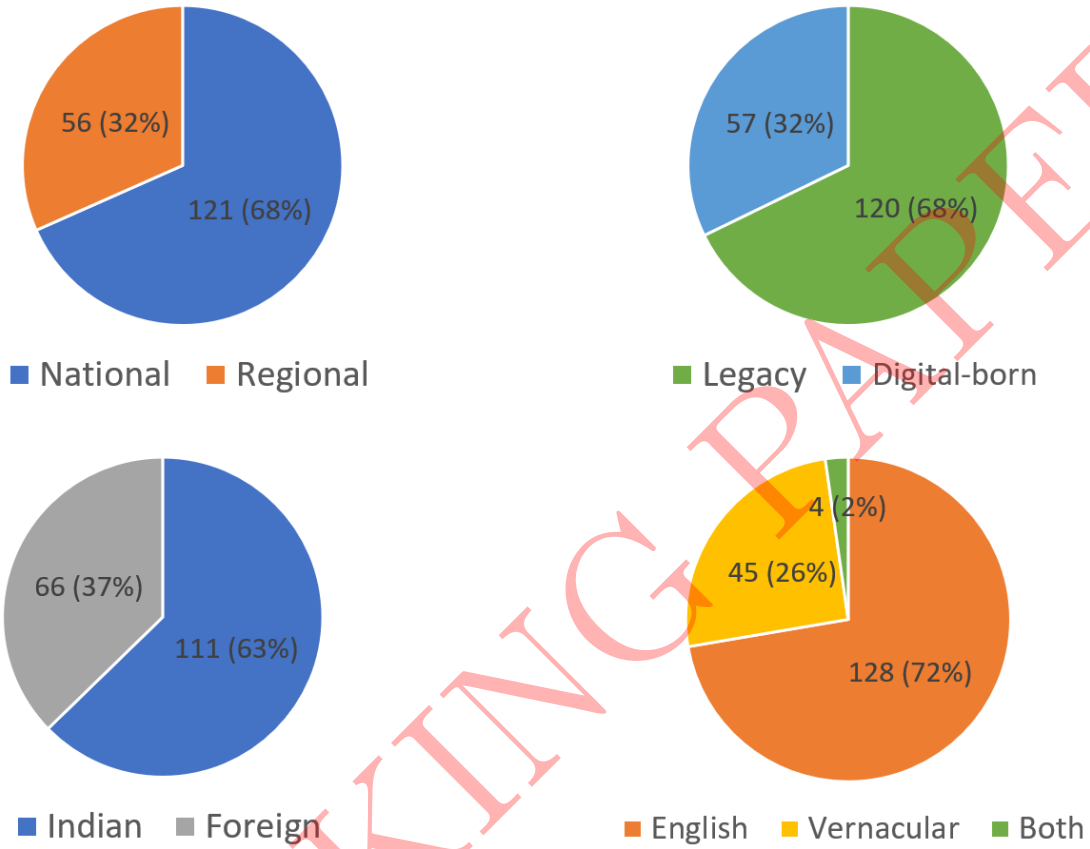

Figure 2: Pie-charts showing the distribution by type of the 177 media outlets that are common to all 45 months.

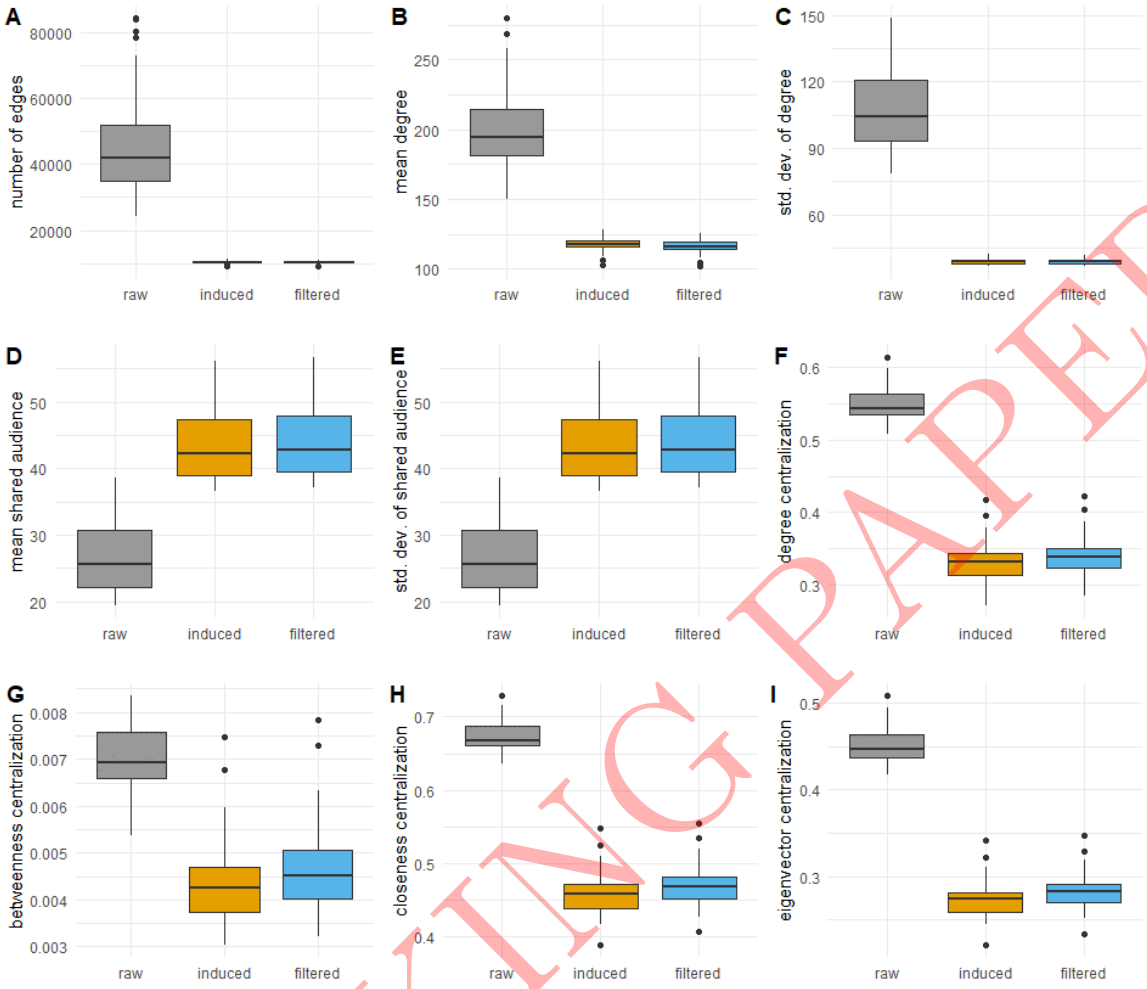

Figure 3: Boxplots showing the distributions of the network properties of the raw networks, induced networks (containing only the common nodes), and filtered networks (after dyadic thresholding): the number of edges (A), mean degree (B), standard deviation of degree (C), mean shared audience or edge weight (D), standard deviation of shared audience or edge weight (E), degree centralization (F), betweenness centralization (G), closeness centralization (H), and eigen-vector centralization (I).

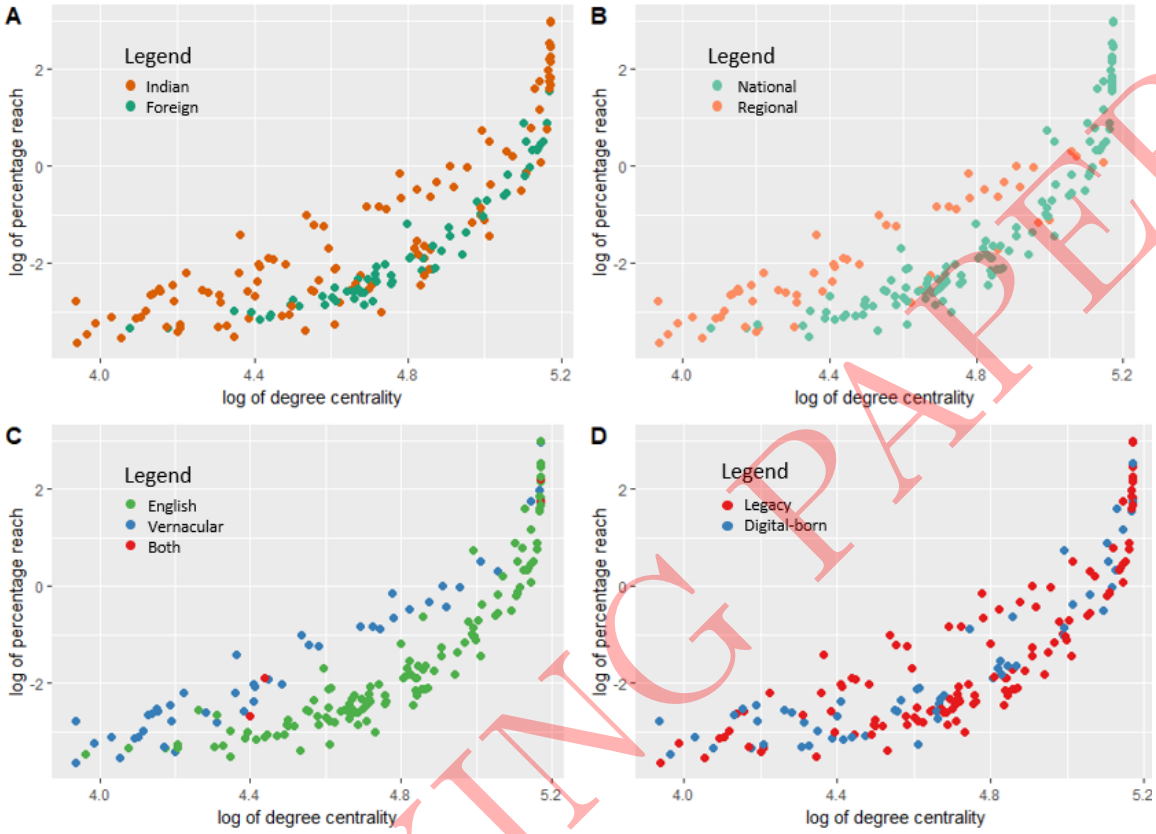

Figure 4: Scatterplots showing the high correlation between percentage reach and degree centrality, colored by type of media outlet. Indian, national, English-language, and legacy media outlets are in general more central and reach greater audiences as compared to non-Indian, regional, vernacular, and digital-born outlets respectively.

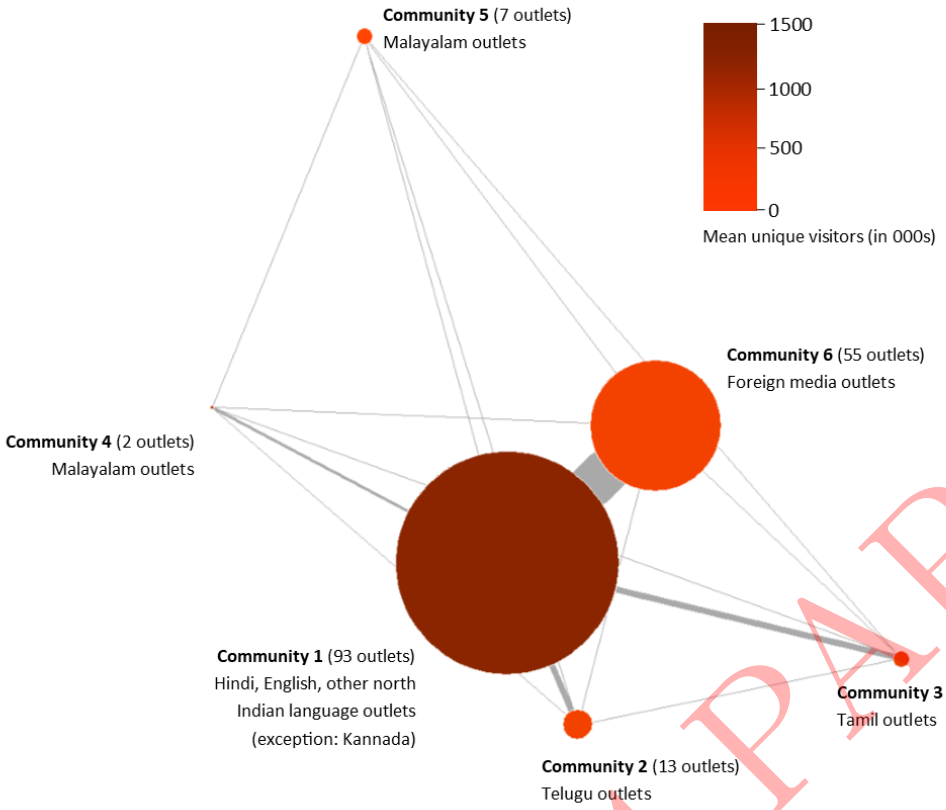

Figure 5: The network of communities obtained by running the WalkTrap algorithm on the network without self-loops showing some linguistic fragmentation of the audience network. The node color shows the average number of unique visitors to media outlet in that community (over the 45-month period). The size of the node shows the number of media outlets included in that community.

5.2 Identification of News Reading Publics

Next, I use the Walktrap algorithm on both, the aggregated network, as well as the augmented aggregated network (i.e. the network with self-loops) and compare findings, to see whether the latter provides a more nuanced set of communities than the former.

Figure 5 shows the results of running the Walktrap community detection algorithm on the aggregated network (i.e. the network obtained by summing the corresponding edge weights for all the monthly networks, but without the self-loops). Despite a slightly negative modularity score (-0.0019), the algorithm is able to identify six distinct communities in the news consumption network.

A breakdown of the communities reveals a clear linguistic divide. Three of these six

communities are linguistically homogenous. Community 2 for instance, contains only Telugu outlets (i.e. the primary language of the states of Andhra Pradesh and Telangana); Community 3 contains only Tamil outlets (corresponding to the state of Tamil Nadu), and Community 4 and 5 contains only Malayali outlets (corresponding to the state of Kerala). Community 1 contains a mix of Hindi, English, and other primarily North Indian languages (Kannada, a South Indian language, being an exception). Community 6, on the other hand is an English-only community, comprising of media outlets that are all international. On the face of it, this partition seems logical. This is because, the majority of Indian languages belong to two linguistic families. Hindi, and other languages that are typically classified as "North Indian" all belong to the Indo-European family and are derived from the classical language called Sanskrit. These languages are similar in that they have all been influenced by each other and are all written in a typical style known as Devnagari, a script that originated in the Indian subcontinent between 100-400 C.E.

One reason why all these North Indian language outlets get sorted into one community is the similarity they share with each other. A second reason is the hegemony of Hindi in the northern states which being amongst the most populated regions on earth, is dominated by a public that speaks Hindi as either a first, second or third language. With the states in the south, the linguistic fragmentation is very different. These states are home to the second major family of languages in India, known as Dravidian. These include Telugu, Tamil, Malayalam, and Kannada. The existence of the Telugu, Tamil, and Malayalam communities in the audience network can then be interpreted through the lens of social identity theory and cultural proximity. In other words, in so far as news reading behavior goes, the Dravidian identity can be identified in the presence of the Telugu, Tamil, and Malayalam communities. The absence of a Kannada community (the other major Dravidian language in India) is surprising, but likely a result of the state of Karnataka being an extremely cosmopolitan state (being the seat of the IT hub Bangalore), drawing people from all across the country with diverse news diets. The identification of an international-only community of outlets also points to the existence of a dedicated international news reading public within the Indian news audience network.

Next, from the network of these six communities (with edge weights signifying the sum of all edge weights between every pair of communities), I find as expected, that the smallest overlaps exist between the Telugu, Tamil, and the two Malayalam communities. The next smallest overlaps exist between these regional communities and the international community, again in line with prior expectations.

Figure 6 shows the results of running the Walktrap community detection algorithm on the augmented aggregated network (i.e. the aggregated network after the addition of self-

loops to every node, with weights equal to the audience reach of the node). The addition of the self-loops not only makes the network more modular (modularity score = 0.005), but also reveals a better partitioning of nodes into fifteen communities. Four of these communities (Communities 12-15) are individual nodes by themselves. This points to them drawing extremely narrow, specialized audiences. Interestingly, a new community of Kannada outlets, hitherto sorted along with other north Indian outlets also emerges. This confirms the suspicions that the news consumption landscape is far more linguistically fragmented in the south, than it is in the north. Moreover, English outlets also get neatly sorted into their own community. Only one community remains that is linguistically heterogeneous, and it contains a mix of Hindi, Marathi, Odiya, and Bengali – all north Indian languages – in line with our expectations.

Observing the network formed by these fifteen communities, I find that the smallest overlaps exist between the linguistically homogenous vernacular communities, while larger overlaps characterize the relationship either between communities of national outlets, international outlets, or between regional outlets and national outlets. The addition of self-loops to the audience network thus demonstrates a positive methodological step for the analysis of audience networks as it helps bring out a more nuanced substructure of how audiences are organized, than what regular community detection without the self-loops allow.

5.3 Longitudinal Trends

Figure 7 shows the longitudinal trends of the percent reach of the 6 different communities identified by the WalkTrap algorithm from the aggregated network. We see that while communities 1 and 6 have seen a slight upward trend, communities 2-5 have seen steeper downward trends over the period from October 2014-June 2018. Community 1 (containing the dominant legacy and North Indian outlets) has understandably been the most popular throughout the period of analysis, and their dominance has seemingly increased over time (significant at the . The international news reading public (Community 6) has also followed an upward trajectory. What this means is that linguistically localized news reading outlets (in Communities 2-5) have lost ground to the dominant legacy and North Indian outlets (in Community 1), and the international outlets (in Community 6).

Figure 8 shows the longitudinal trends of the percent reach of the 11 different non-singular communities identified by the WalkTrap algorithm from the augmented aggregated network and helps clarify the comparison between the temporal trends further. As before,

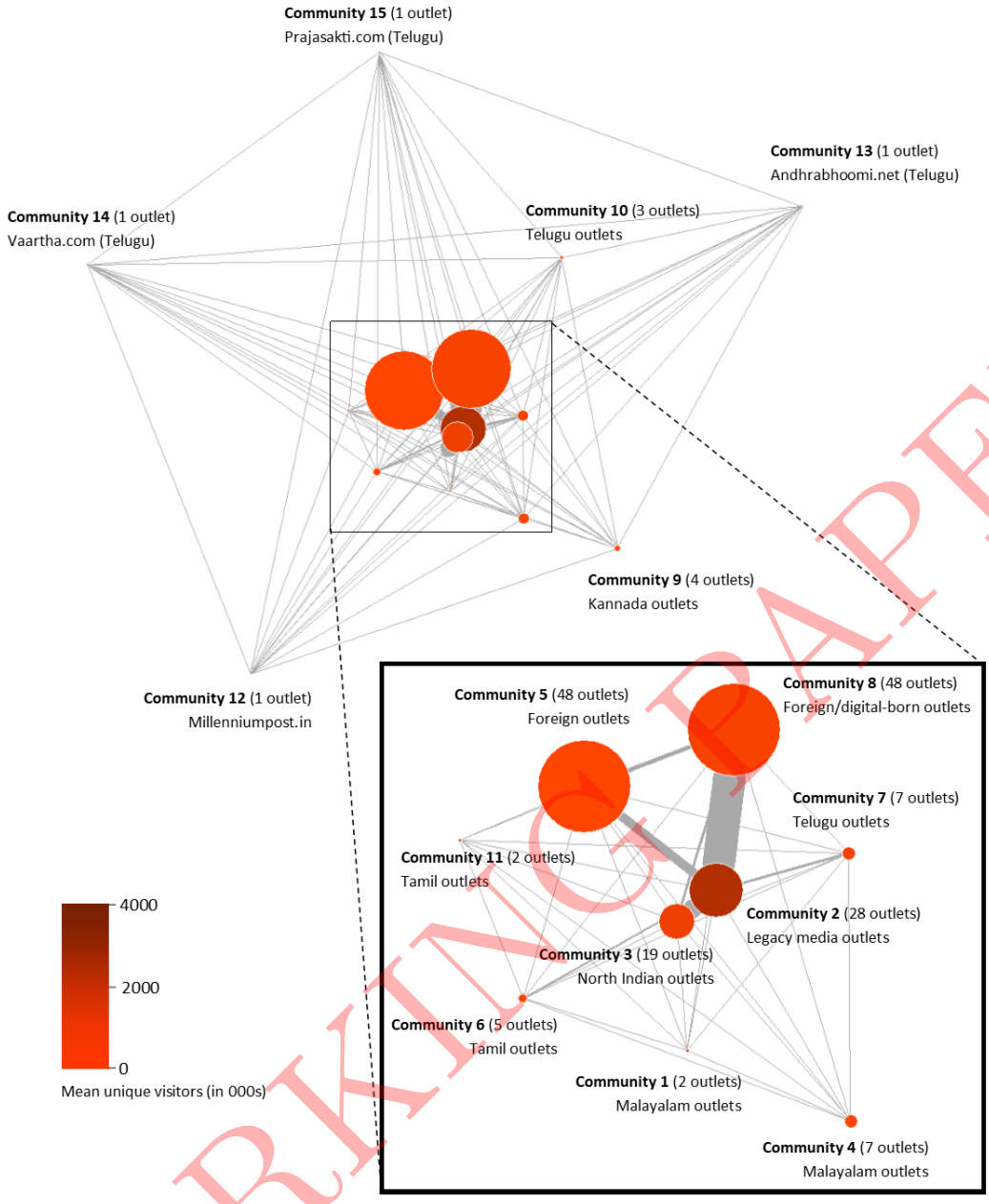

Figure 6: The network of communities obtained by running the WalkTrap algorithm on the network with self-loops shows clearer extent of linguistic fragmentation of the audience network. The node color shows the average number of unique visitors to media outlet in that community (over the 45-month period). The size of the node shows the number of media outlets included in that community. The core is shown separately magnified.

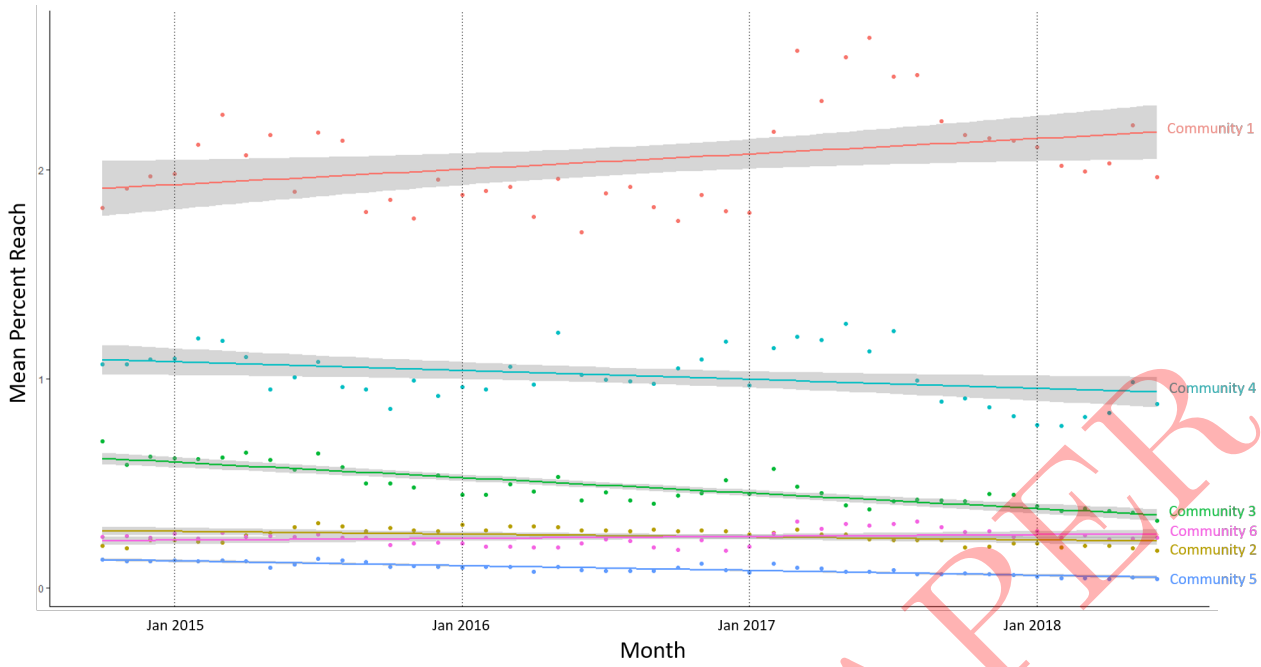

Figure 7: Temporal trends show that while Communities 1 (national legacy outlets) and 6 (international outlets) have witnessed a significant increase in audience reach, Communities 2-5 (all vernacular outlets) have lost out.

the only publics to witness a statistically significant increase in percent reach over time were the legacy media (Community 2) and international media (Community 5) reading publics respectively. All non-singular communities, corresponding to vernacular languages (Communities 1; 3-7; 9-11) have seen statistically significant declines over the period of analysis.

This raises concerns about the health of vernacular media in India, which appear to have been left behind amid the staggering growth of the national legacy media, and the prevailing globalization that has made more and more Indians look to international outlets for news.

6 Discussion

From a theoretical perspective, news reading publics exist in all countries, in various different contexts. The case of India is more useful than unique in this regard. The decision to focus

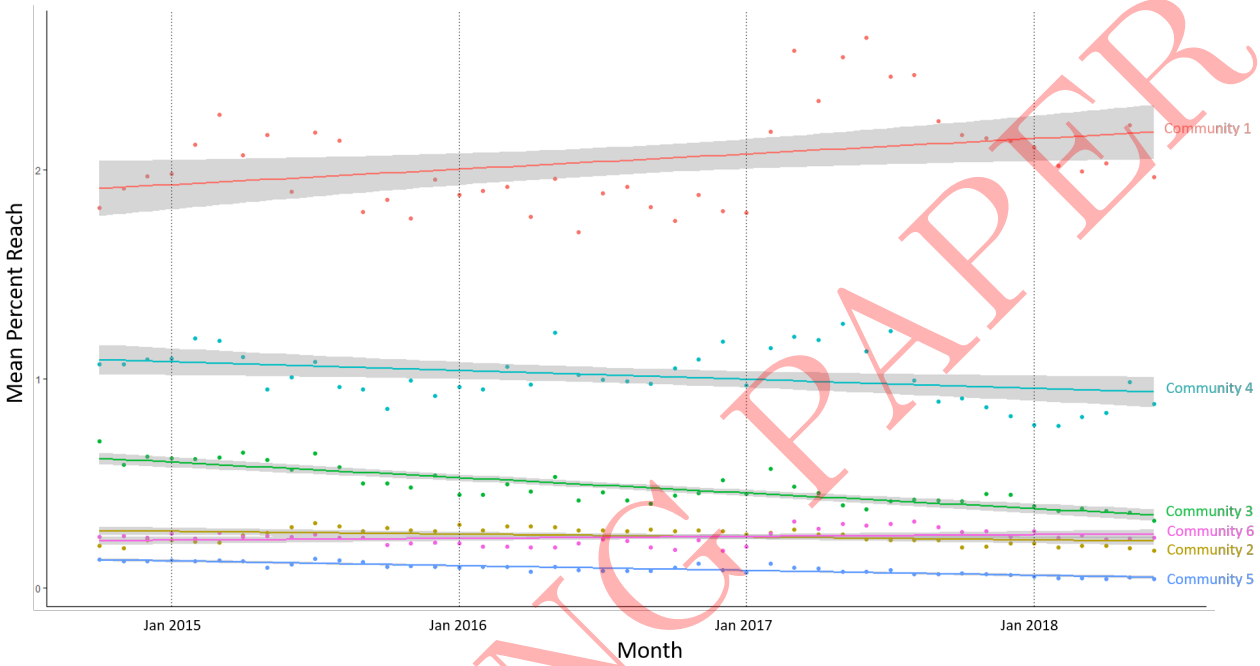

Figure 8: Temporal trends show that while Communities 2 (legacy English national outlets) and 5 (international outlets) have witnessed a significant increase in audience reach, Communities 1, 3-7 and 9-11 (each of which corresponds to a specific regional news reading public) have lost out. The slope of Community 8 is 0. Trend lines of communities 12-15 are not shown as they contain single outlets.

on India for informing this theory is driven by its cultural and linguistic heterogeneity, that makes it possible to identify, operationalize and quantitatively define news reading publics from audience data. As has been explained before, this is primarily because language serves as a useful marker to categorize the news media at the outlet level, and also serves as a proxy for shared culture, identity, and gratifications that audiences expect out of consuming news. With the US, the UK or any other country which has lesser linguistic heterogeneity, one would either need more granular data to identify these publics, as there is no such clear marker that unites the production and consumption of news. Thus, news reading publics in these countries are typically characterized along ideological or demographic lines. Selective consumption and avoidance of political news based on partisan identity, or the generational or demographic specialization of media outlets then become two special cases of the theory of news reading publics – i.e, they describe the news reading habits of publics that are either ideologically homogenous, or demographically homogenous.

One advantage of using India to inform this framework is that it increases the external validity of the theory. If we think of cultural heterogeneity to be a continuum, with India at one end, then we can begin to think of how news consumption patterns can look like in democracies that are not as heterogenous, given that we already understand what drives news reading behavior in a country like the US. Spain, for instance, has five different languages, three of which (Spanish (Castilian), Catalan, and Galician) spoken by significantly large numbers of people as a first or second language (99%, 19%, and 5% respectively). Knowing how the audience overlap network looks like in the case of India, we can expect a similar (although to a lesser extent) modular structure in the case of Spain. Using evidence from multiple different contexts such as these, we can contribute towards making this framework richer, thereby shedding light on the extent to which different shared commonalities within an audience determine their news consumption habits in different countries.

The identification of news reading publics thus demonstrates an important theoretical step for understanding commonalities in news consumption behavior. Moreover, the it also lets us trace their relative evolutions over time, allowing us to compare their temporal trends, and assessing the health of the news consumption landscape in general. In the case of India, I find evidence of the increasing hegemony of the legacy national media over vernacular media, and two, an increasingly international outlook among the general news reading public. While these trends are possibly attributable to the changing socio-economic demographics of a fast developing country (particularly the shift from a rural agrarian to an urban industrial economy, and the rising literacy rate), they raise questions over what the future holds for vernacular media, which still continue to dominate the broadcast and print formats.

As with any research however, there are some limitations of the study described in this paper. Perhaps the most salient limitation is the fact that the data I use are only available in monthly aggregated formats. It is therefore impossible to analyze the data at a more fine-grained or individual level. This in turn, makes it impossible to distinguish between users based on their visit frequency or duration. Studies that achieve to do that by tracking the individual level activity [2, 13] tend to compromise on external validity by sampling from non-representative populations (for instance, Microsoft Bing users). While the use of online panel data like Comscore's helps circumvent some of those issues (since they claim to maintain representative panels of users in the various countries they operate in) it raises concerns about transparency. As has been described earlier, Comscore uses a proprietary algorithm to generate its monthly aggregates by integrating server-side data with individual browsing data obtained from a piece of installed software on their panelists' devices. The manner in which the integration is done remains unknown. However, the analysis of digital trace data of web browsing behavior, has several documented advantages when compared to self-reports of news and information exposure [11, 16, 39]. While the former is not free of its own systemic biases, the latter suffer from other kinds of survey errors [17, 50], including but not limited to incorrect recall. The use of trace data, to that end, offers a relatively more "accurate representation of how people consume news and how those patterns change over time and across media environments." [32].

The final limitation of this study relates to the fact that the dataset I use only captures desktop browsing activity. While it is true that mobile phones have witnessed phenomenal high growth and surge in popularity in India in recent years, there is some evidence to suggest that only a small fraction of mobile phone users in India use their phones to read news [33]. Moreover, even if hypothetically we assume that a majority of mobile phone users do consume news on their phones, I would expect to find greater evidence of more distinct news reading publics in the network. In fact, a few of the more socioeconomically backward regions that aren't seen in my network would likely appear if mobile browsing data was used instead. Finally, it is important to view the Comscore data from the perspective of existing research endeavor. All virtue is relative and given the scant empirical evidence that is available about news consumption in India today, Comscore remains the most comprehensive yet accessible source of reliable audience data.

References

[1] Allcott, H., and Gentzkow, M. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives 31*, 2 (2017), 211–236.

[2] Athey, S., Mobius, M., and Pal, J. The Impact of Aggregators on Internet News Consumption.

[3] Bauer, R. A. The Obstinate Audience. *The Process and Effects of Mass Communication*, April (1971), 326–346.

[4] Blumler, J. G., and Katz, E. *The Uses of Mass Communications: Current Perspectives on Gratifications Research. Sage Annual Reviews of Communication Research Volume III.* Sage Publications, Inc., 275 South Beverly Drive, Beverly Hills, California 90212 ($17.50 cloth, $7.50 paper), 1974.

[5] Boyd, D. A., and Najai, A. M. Adolescent TV Viewing in Saudi Arabia. *Journalism Quarterly 61*, 2 (6 1984), 295–351.

[6] Chadha, K., and Kavoori, A. Media imperialism revisited: some findings from the Asian case. *Media, Culture & Society 22*, 4 (2000), 415–432.

[7] Chakravartty, P., and Roy, S. Media Pluralism Redux: Towards New Frameworks of Comparative Media Studies "Beyond the West". *Political Communication 30*, 3 (2013), 349–370.

[8] de Sola Pool, I. The Changing Flow of Television. *Journal of Communication 27*, 2 (6 1977), 139–149.

[9] Del Vicario, M., Zollo, F., Caldarelli, G., Scala, A., and Quattrociocchi, W. Mapping social dynamics on Facebook: The Brexit debate. *Social Networks 50* (2017), 6–16.

[10] Delli Carpini, M. X., and Keeter, S. *What Americans know about politics and why it matters.* Yale University Press, 1996.

[11] Dilliplane, S., Goldman, S. K., and Mutz, D. C. Televised Exposure to Politics: New Measures for a Fragmented Media Environment. *American Journal of Political Science 57*, 1 (2013), 236–248.

[12] Fiorina, M. P., Abrams, S. J., and Pope, J. C. *Culture War? The Myth of a Polarized America.* 2006.

[13] Flaxman, S., Goel, S., and Rao, J. M. Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly 80*, Specialissue1 (2016), 298–320.

[14] Garrett, R. K., and Stroud, N. J. Partisan Paths to Exposure Diversity: Differences in Pro- and Counterattitudinal News Consumption. *Journal of Communication 64*, 4 (8 2014), 680–701.

[15] Giddens, A. *The constitution of society-outline of the theory of structuration*. Polity Press, 1984.

[16] Goldman, S. K., Mutz, D. C., and Dilliplane, S. All Virtue Is Relative: A Response to Prior. *Political Communication 30*, 4 (2013), 635–653.

[17] Groves, R. M., and Lyberg, L. Total Survey Error: Past, Present, and Future. *Public Opinion Quarterly 74*, 5 (1 2010), 849–879.

[18] Heider, F. *The psychology of interpersonal relations*. Erlbaum, 1958.

[19] Hogg, M. A., Terry, D. J., and White, K. M. A Tale of Two Theories : A Critical Comparison of Identity Theory with Social Identity Theory. *Social Psychology Quarterly 58*, 4 (1995), 255–269.

[20] Kang, J. G., and Morgan, M. Culture Clash: Impact of U.S. Television in Korea. *Journalism Quarterly 65*, 2 (6 1988), 431–438.

[21] Katz, E. Mass Communications Research and the Study of Popular Culture : An Editorial Note on a Possible Future for This Journal Mass Communications Research and the Study of Popular Culture : An. *Studies in Public Communication 2* (1959), 1–6.

[22] Kelley Harold H. *Attribution in Social Interaction*. 1971.

[23] Kumar, S. Media Industries in India: An Emerging Regional Framework. *Media Industries 2* (2014).

[24] Lancichinetti, A., and Fortunato, S. Community detection algorithms: A comparative analysis. *Physical Review E 80*, 5 (2009), 056117.

[25] Lancichinetti, A., and Fortunato, S. Erratum: Community detection algorithms: A comparative analysis (Physical Review E - Statistical, Nonlinear, and Soft Matter Physics (2009) 80 (056117)). *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics 89*, 4 (2014), 56117.

[26] Levendusky, M. Partisan Media Exposure and Attitudes Toward the Opposition. *Political Communication 30*, 4 (2013), 565–581.

[27] Luhmann, N. *The reality of the mass media*. Stanford University Press, 2000.

[28] Majó-Vázquez, S., Mukerjee, S., Neyazi, T. A., and Nielsen, R. K. Online Audience Engagement with Legacy and Digital-Born News Media in the 2019 Indian Elections. Tech. Rep. May, Reuters Institute for the Study of Journalism, Oxford, UK, 2019.

[29] McCall, G. J., and Simmons, J. L. *Identities and interactions*. Free Press, New York, 1978.

[30] Mead, G. Mind, Self, and Society: From the Standpoint of a Social Behaviorist, Mead, Morris, 1934.

[31] Messing, S., and Westwood, S. J. Selective Exposure in the Age of Social Media: Endorsements Trump Partisan Source Affiliation When Selecting News Online. *Communication Research 41*, 8 (2014).

[32] Mukerjee, S., Majó-Vázquez, S., and González-Bailón, S. Networks of Audience Overlap in the Consumption of Digital news. *Journal of Communication 68*, 1 (2018), 26–50.

[33] Nayak, D. Mobile Content Consumption In India Is Primary Driven By Entertainment & Sports [REPORT] - Dazeinfo, 2018.

[34] Peterson, E., Goel, S., and Iyengar, S. Echo Chambers and Partisan Polarization : Evidence from the 2016 Presidential Campaign.

[35] Pons, P., and Latapy, M. Computing Communities in Large Networks Using Random Walks Pascal. *Journal of Graph Algorithms and Applications 10*, 2 (2006), 191–218.

[36] Price, V. Conceptualizing the Public. In *Public Opinion*. Sage Publications, London, 1992, pp. 22–44.

[37] Price, V., David, C., Goldthorpe, B., Roth, M. M. C., and Cappella, J. N. Locating the issue public: The multi-dimensional nature of engagement with health care reform. *Political Behavior 28*, 1 (2006), 33–63.

[38] Prior, M. *Post-broadcast democracy : how media choice increases inequality in political involvement and polarizes elections*. Cambridge University Press, 2007.

[39] Prior, M. The Challenge of Measuring Media Exposure: Reply to Dilliplane, Goldman, and Mutz. *Political Communication 30*, 4 (2013), 620–634.

[40] Rajagopal, A. *Politics after Television: Hindu Nationalism and the Reshaping of the Public in India*. Cambridge University Press, 2004.

[41] Salwen, M. B. Cultural Imperialism: A Media Effects Approach. *Critical Studies in Mass Communication 8* (1991), 29–38.

[42] Stryker, S. Identity Salience and Role Performance: The Relevance of Symbolic Interaction Theory for Family Research. *Journal of Marriage and the Family* (1968).

[43] Stryker, S. Identity theory: Developments and extensions. In *Self and identity: Psychosocial perspectives*. John Wiley & Sons, Oxford, England, 1987, pp. 89–103.

[44] Stryker, S., and Serpe, R. T. Commitment, Identity Salience, and Role Behavior: Theory and Research Example. In *Personality, Roles, and Social Behavior*. 1982.

[45] Stryker, S., and Statham, A. Symbolic Interaction and Role Theory. In *Symbolic Interactionism*. 1977.

[46] Tan, A. S., Tan, G. K., and Tan, A. S. American TV in the Philippines: A Test of Cultural Impact. *Journalism Quarterly 64*, 1 (3 1987), 65–144.

[47] Thompson, J. B. J. B. *The media and modernity : a social theory of the media*. Stanford University Press, 1995.

[48] Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., and Wetherell, M. S. *Rediscovering the social group: A self-categorization theory*. Blackwell, Oxford, 1987.

[49] Weiner, B. Attribution theory, achievement motivation, and educational process. *Review of Educational Research 42*, 2 (1972), 203–215.

[50] Weisberg, H. F. *The Total Survey Error Approach: A guide to the new science of survey research*. University of Chicago Press, 2005.