Defining Local News: A Computational Approach

Nick Hagar
nicholashagar2018@u.northwestern.edu
Northwestern University

Jack Bandy
jackbandy@u.northwestern.edu
Northwestern University

Daniel Trielli
dtrielli@u.northwestern.edu
Northwestern University

Yixue Wang
yixue.wang@u.northwestern.edu
Northwestern University

Nicholas Diakopoulos
nad@northwestern.edu
Northwestern University

Abstract

This paper presents a computational method for defining local news outlets. The decline of local journalism has prompted widespread research efforts, but these efforts are currently hampered by muddled and conflicting definitions of local news. To help gather efforts around a single definition, our method utilizes openly-available volunteered geographic information from Twitter users who follow news outlets. We find distinct geographic signals between the audiences of local, regional, and national news outlets. We also find that the *in-state audience rate* differs between these classifications. Moving forward, this audience-based method can help scholars and practitioners accurately identify local news outlets – a crucial step to providing meaningful support.

ACM Reference Format:
Nick Hagar, Jack Bandy, Daniel Trielli, Yixue Wang, and Nicholas Diakopoulos. 2019. Defining Local News: A Computational Approach. In *Computational + Journalism Symposium 2020*. ACM, New York, NY, USA, 5 pages.

1 Introduction

Local news organizations play a demonstrable civic role in their communities. Community newspapers place their readers in the context of the larger world [14], support collective identities, provide opportunities to discuss community issues [18], and more. This position impacts tangible economic and political facets of a community, from municipal borrowing costs [10] to participation in local elections [22, 23].

But for all the importance of local news organizations, as well as efforts to document their decline [1] and prevent further losses [3], there is surprisingly no consistently used and computationally scalable definition in existing literature: *What exactly constitutes a local news organization?*

Researchers have previously used definitions of "local" that allow quick, straightforward classifications of news publishers. On the print side, several definitions use a circulation of 50,000 as the cut-off point for "small-market papers" or "community papers" [12, 21]. Lauterer (as cited in [12]) defines these as papers which "serve people who live together in a distinct geographical space with a clear local-first emphasis on news, features[,] sport and advertising" (p. 49-50). Another study of local news relied on lists of U.S. newspapers maintained by a trade group [10]. On the digital side, Hindman constructs a sample of local news outlets with the following definition: "sites that have higher levels of usage within a given media market than they do in the rest of the nationwide sample" [15, p. 4]. Coverage has also been used as an important identifier of local news [19].

In this work, we take initial steps toward developing a computational definition of local news. Rather than relying on audience size heuristics or topical selections, we analyze *geographic* patterns of news audiences. We approach this work from two perspectives. First, we manually code 6 news outlets as local, regional, or national based on the geographic reach of their print newspapers. As a broader exploration, we also build a larger-scale (N=100) dataset of news outlets generated from Comscore. In both datasets, we find that audiences of local news outlets (as measured by the outlet's Twitter followers) exhibit distinct geographic patterns. Specifically, our analysis shows (1) that local news audiences are distinctly clustered around the location of the newsroom, and (2) that the *in-state audience rate* is substantially higher for local outlets than for regional and national outlets.

2 Background

The measures used to distinguish local news from other kinds of journalism in past work broadly rely on three characteristics of a news organization: *coverage*, *production*, or *audience*. Across these groups, there is no concrete, widely-accepted definition for what makes a news outlet local. In fact, Lowrey et al. [18] analyzed 108 studies on local news and found that only 30 offered any explicit definition of the term. Many frameworks offer general conceptual guidelines for determining news "localness," but a handful of researchers have also created specific, formal operationalizations.

Coverage

Ali [2] contextualizes local news coverage as either place-based or content-based. Geography and proximity play important roles in place-based coverage, as outlets cover the

C+J '20, March 20–21, 2020, Boston, MA
.

events that happen in their community [4]. The idea of geographic proximity is a common concern across prior work, with references to "regionally specific news" [16], "proximity" coverage as a news value [14], and election coverage "that affects local people" [7]. This perspective provides a straightforward view of the content of local news coverage as well–anything that concerns the everyday matters of a community [17]. Other work has identified crime as the leading topic for local television news coverage [9]. Place-based locality of coverage has also been measured by analyzing the geography of front-page stories [5].

Content-based local coverage is less well-defined, as it can refer to anything that a community finds interesting [2, 16]. This can mean focusing on local stories by local journalists, or it can mean presenting a mix of regional, national, and international coverage [7, 16, 18, 19]. This perspective is subject to audience demand: Locally-focused coverage has previously been observed to increase over time, as readers seek it out more [7]. Other work also points to the importance of news coverage itself in defining locality [14, 18]. In this view, local news coverage is that which both reflects and helps create the common view of a community.

Production

In this context, production refers to the outlets that publish news. In broadcast journalism, a news outlet must be located and invested in a community to meet the licensing requirements of a local station [2]. Print media has fuzzier guidelines. Researchers have pointed to smaller staff sizes or commercial operations as indicators of a local news outlet, or to the distribution schedule of the newspaper [5, 8, 13, 14, 16, 17].

Others attempt to define local news production by its explicit association with a place. Hess and Waller [14] note that small, commercial newspapers tied to specific communities are often identified by a place name in their masthead. Other researchers label local news by the relative geographic size and density of their communities, referring to "suburban papers" or "provincial non-daily newspapers" [17]. In all cases, this perspective asserts that some trait connects a local news outlet with a geographic community.

Audience

In the most restrictive sense, a local audience is one that is located in a specific geographic area (a neighborhood or town) [4]. However, many communities fall outside of that definition, especially since communities can emerge from shared meaning – not merely shared location [2, 18]. Also, as readers' interests and concerns shift, so does the content they might consider local [16]. And in fact, the "localness" of content often has more to do with the news *topic* than with geography [19]. Finally, while some may care about everyday events in a community, not all news in a geographic area is equally salient to all readers [17, 19].

Even within geographical bounds, the idea and importance of local varies. Immigrants move from place to place, complicating the idea of "local" as a single location [4]. Communities can include neighborhoods and cities, but they can also define ethnic, online, or shared-interest groups [18]. A people-oriented notion of local allows members of a community to feel a "sense of being" without being located in a particular place [4]. To encapsulate these new kinds of communities, some researchers have begun referring to "geo-social" journalism as the news that applies to certain connected places and social networks [12, 14]. An audience view, then, suggests a more expansive definition of local news.

Toward a formal definition

We aim to develop a straightforward, scalable, computational method for assessing the localness of digital news outlets using openly-accessible data. Many definitions explored in this section do not meet these goals because they are medium-specific. For example, print newspaper circulation cannot be applied to digital outlets. Other definitions, such as that used by Hindman [15], rely on proprietary data.

This study addresses these challenges by introducing a computational heuristic for outlet localness that is reproducible and scalable. Because this method incorporates publicly available data to classify news organizations based on geographic scope, other researchers can reuse it with new data, whether to expand the coverage to other news organizations not in our sample, to use the method to support their own analyses, or to validate the method against other available data (proprietary or otherwise).

While one can imagine a computational approach to all of the areas explored above, we focus on an audience location perspective. This is because audience analysis is more expansive than other approaches (e.g., content analysis or economic impact), and it provides a viable proof of concept that can be rapidly applied in future work. In the next section, we discuss the methods we developed and deployed.

3 Methods and Data

Our overall approach involves analyzing geographic patterns in a news outlet's Twitter audience. Specifically, it consists of (1) matching news outlets to Twitter accounts, (2) collecting IDs of users who follow a news account on Twitter, (3) collecting the location field from a statistically significant sample of those users' profiles, (4) geocoding the location fields into geographic coordinates, and (5) analyzing geographic patterns in the geocoded locations.

Twitter Data

Outlet Selection and Twitter Matching We use two datasets in our analysis. The first is a hand-selected set of news sources consisting of two national, two regional, and two local news outlets (anonymized for submission):

- New York Times (national)
- USA Today (national)
- Chicago Tribune (regional)
- Chicago Sun-Times (regional)
- Evanston Now (local)
- Evanston Roundtable (local)

While the first set of sources was intentionally small, to allow for manual checking of their historical geographic reach, a second set was used for more extensive evaluation. It consists of local news sources sampled from the top websites in Comscore's "Local News" category (also utilized by Hindman [15]). We scraped each local news website to check for a Twitter account linked on the home page, which matched 445 Twitter accounts to local news outlets. Due to Twitter's rate-limited API, we used a random selection of one hundred of these accounts in our analysis.

Collecting Location Information Our audience location data uses the "location" field of Twitter user profiles. Collecting location information for all followers would have taken several months through Twitter's API, so we devised a sampling method for each of the two datasets, recognizing that not all users enter valid geographic information in this field [11].

To calculate the appropriate sample size of followers for the six manually-chosen outlets, we calculated an appropriate sample size for a 99% confidence level and a 1% margin of error on the self-reported Twitter data for the outlet with the highest number of followers among the six – The New York Times, with 44.5 million followers. For that population of followers, the minimum appropriate sample size was 16,635 geocoded followers. We thus aimed to collect 17,000 valid locations per outlet, and *all* user locations for local news sources since they had fewer than 17,000 total followers.

Since many Twitter users leave location information blank, we expected the geocoding process (described in the next subsection) to yield a substantially smaller sample of valid locations compared to the total number of users inputted. Conservative estimates from preliminary data suggested that 1 in 5 location fields from Twitter would yield a valid location. Therefore, to reach 17,000 valid follower locations per news organization, we gathered data for 100,000 random followers of each news organization (except the two local outlets).

For the second dataset comprising 100 local news outlets, we targeted a proportion estimate with 95% confidence and a 5% margin of error, which required a minimum of 385 valid locations. Based on results from geocoded locations for the smaller dataset, we estimated that approximately 1 in 3 users would list their location; to reach our desired confidence level, we thus collected a random sample of 1,500 followers for outlets with more than 1,500 total followers.

Geospatial Data

Geocoding Across the six manually-selected publications, the proportion of users who listed any location on their profile ranged from 29% to 66%. We converted the "location" field from these Twitter profiles to a geographic location using ArcGIS [6]. This process turns the location string from a user's profile into a location name and associated latitude and longitude coordinates. It is important to note that our geocoding process only matches locations within the United States, putting a geographic upper bound on our results. We undertake two filtering steps to limit our sample to valid locations. First, when ArcGIS geocodes a location, it generates a confidence score ranging from 0 to 100. We only include profiles with a score of 100. Second, we filter remaining locations to only those that include a comma. This step helps ensure that we only capture locations that follow a "city, state" formulation. While ArcGIS can return state names on their own as valid geocodes, these locations are too broad to accurately place on a latitude/longitude point (e.g. "New York, NY" versus "New York"). Across outlets, 17% to 100% of followers with a geocoded location match these additional constraints. Based on a manual check of 1,000 user profiles, this pipeline produces an error rate of 0.5%.

Analysis We take two steps to demonstrate the geographic differences between audiences of local, regional, and national outlets. First, we measure the most common city and most common state among an outlet's followers, and the percentage of followers in that city and state. We refer to these percentages as the *in-city audience rate* and the *in-state audience rate*. Second, we calculate the distance, in miles, from each follower to the news outlet, to create *follower distance distributions*. We then look for evidence of differences between the local, regional, and national outlets in their distributions via cumulative distribution plot.

4 Results

Results show clear distinctions between audience locations for local, regional, and national news outlets. Table 1 shows a clear difference in the percent of followers located in the same state as each news outlet. For every group, we see a much smaller difference in percentage between outlets *within* a group (mean=6%) than from one group to the next largest (mean=35%).

Figure 1 further demonstrates these differences in a more granular view. It shows the cumulative distribution of follower distances for each outlet (e.g., approximately 75% of the *New York Times*' followers are within 1,500 miles of the paper). From this view, we see two patterns. First, the distributions of outlets *within* each group (e.g., Local 1 to Local 2) are closer together than to those outside the group. Second, the distributions within each grouping tend to follow similar trends. National outlets' followers show the most geographic

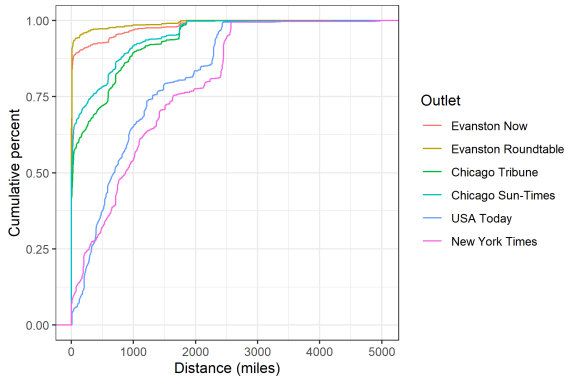

Figure 1: Cumulative percentages for each outlet's follower distance distribution

spread and are the most uniformly dispersed across the distribution. Regional outlets have some geographic variation, with many followers relatively close by or in the same city. Finally, local outlets' followers are almost all in or near the same city as the outlet.

Outlet	N	Top City	% In City	Top State	% In State
EvRoundTable	654	Evanston IL	63.30%	IL	95.72%
EvanstonNow	2405	Evanston IL	44.86%	IL	89.94%
chicagotribune	23698	Chicago IL	41.92%	IL	59.53%
Suntimes	25701	Chicago IL	50.38%	IL	67.61%
USATODAY	20616	Washington DC	3.28%	CA	11.02%
nytimes	29525	New York NY	4.50%	CA	14.61%

Table 1: Descriptive statistics for the locations of each outlet's Twitter followers

One Hundred Random Outlets

In-state audience rate remains a strong indicator of local news in the 100 random outlets from Comscore's local news category, with in-state audience rates above 70% for 89 of the outlets. The rates ranged from 50.70% to 93.37%, with a median of 83.29%. The results also show that some outlets in Comscore's list of "local news" websites are false positives, as noted in Hindman's 2011 report [15]. For example, the Twitter accounts "sedgwickcounty" and "countyofsb" (from https://www.sedgwickcounty.org/ and http://countyofsb.org/) are not local news outlets, but government accounts from Sedgwick County, Kansas, and Santa Barbara, California. In fact, the former was one of the six accounts with the highest in-state audience rate, as seen in table 2.

Measuring the in-state audience rate also surfaced outlets in Comscore's local news category that cover metropolitan (ex. wsbtv), state-wide (ex. UnionLeader), and multi-state (ex. mynbc5) news. We can thus imagine in-state audience rate as a metric for filtering out non-local news outlets from an initial set of sources. The six outlets with the lowest in-state audience rate in the dataset are shown in table 3.

USER_following_outlet	N	top_city	pct_in_city_s	top_state	pct_in_state_s
ksatnews	362	San Antonio TX	70.99%	TX	93.37%
abc30	319	Fresno CA	39.5%	CA	92.79%
ivdailybulletin	437	Los Angeles CA	9.84%	CA	92.68%
SGVTribune	299	Los Angeles CA	25.75%	CA	91.3%
KTVU	236	San Francisco CA	18.22%	CA	91.1%
sedgwickcounty	695	Wichita KS	72.09%	KS	91.08%

Table 2: The six outlets from the Comscore-based dataset with the *highest* in-state follower rate, surfacing outlets that serve a distinctly local audience. The "sedgwickcounty" Twitter account is a government website that Comscore included in its list of "local news" domains.

Outlet	N	Top City	% In City	Top State	% In State
mynbc5	357	Burlington VT	17.65%	VT	50.7%
thecolumbian	420	Vancouver WA	38.33%	WA	56.67%
MauiNow	420	Maui HI	21.9%	HI	59.05%
NBC6News	414	Shreveport LA	32.13%	LA	61.11%
wsbtv	243	Atlanta GA	33.74%	GA	62.96%
UnionLeader	354	Manchester NH	13.56%	NH	63.28%

Table 3: The six outlets from the Comscore-based dataset with the *lowest* in-state follower rate, revealing that some outlets in Comscore's "local news" category are metropolitan (ex. wsbtv), state-wide (ex. UnionLeader), and multi-state (ex. mynbc5) news outlets.

5 Discussion

In this work, we have taken the first step in a larger goal to develop a computational definition of local news. Our analysis shows that it is possible to distinguish a local news publisher from a regional or national publisher based on the location of its Twitter followers.

We conceive this as a first step in a larger construction of a computational method of defining local news. For this to happen, additional future work will be necessary.

First, Twitter is not representative of the entire universe of news consumption [20, 24], and we thus plan to expand our method to other platforms. As we expand the scope of our analysis to other news consumption venues, we anticipate specific challenges for each venue, as well as the overall challenge of combining all analyses into one framework. For example, it is possible that a news organization would be classified as local on Twitter, but not on Facebook. It will be interesting to explore distinctions of localness between platforms, and the implications of such distinctions.

Second, while we start with the analysis of audiences, we also envision methods to classify news *content* as local or not, through text analysis of a news organization's publications. For example, such a method could account for a report from the New York Times about voters in rural Iowa. Results from a content-centric analysis may contrast with the audience-centric definition presented in this paper. In fact, we expect some contrasts—some news organizations may reach local audiences, while others produce local content. Examining

the distribution of audiences and content may therefore lead to new conceptual insights that help distinguish different types of news outlets.

Third, we hope to address a couple limitations of our current approach. As noted in previous literature, an audience view of locality expands beyond geography alone. This study takes the most restrictive perspective of a local audience, equating geographic proximity with locality. Using additional data, such as follower interactions on Twitter or publication readership, we may be able to produce a more nuanced picture of the relationship between the location and engagement of an outlet's audience. In addition, despite our random sampling approach, there is a potential threat to validity in the fact that Twitter users supply their own location information. There may be unobserved qualitative differences between news outlet followers who do and do not provide a location. Bias may also arise in the geocoding stage—the geocoder we use here may interpret some locations more accurately than others, and as noted, it does not account for international locations. In the future, we aim to qualitatively examine potential user differences, as well as compare results from multiple geocoding tools.

Finally, we aim to validate our computational definition using qualitative methods (e.g., surveys and interviews). It is critical to evaluate the continuity between an organization's self-identity, its identity as perceived by other practitioners, and its identity as ascertained by computational heuristics.

By incorporating audience location data from other platforms, examining an outlet's content, and expanding our qualitative validation, our method can help illuminate what it means to be a local news organization. We envision this work to have deep practical implications, such as creating a clear, reproducible, and scalable database of local news organizations. We hope to eventually publish such a database for any person or organization, such as scholars who want to study this crucial subset of the industry or institutions that want to provide support.

References

[1] Penelope Muse Abernathy. 2018. *The expanding news desert.* University of North Carolina Press Chapel Hill.

[2] Christopher Ali. 2017. *Media Localism: The Policies of Place.* University of Illinois Press.

[3] Christopher Ali and Radcliffe Damin. 2017. 8 Strategies for Saving Local Newsrooms. *Columbia Journalism Review* (2017).

[4] Hau Ling Cheng. 2005. Constructing a Transnational, Multilocal Sense of Belonging: An Analysis of Ming Pao (West Canadian Edition). *Journal of Communication Inquiry* 29, 2 (2005), 141–159. https://doi.org/10.1177/0196859904273194

[5] Rick Edmonds. 2007. The 'Local-Local' Strategy: Sense and Nonsense. https://niemanreports.org/articles/the-local-local-strategy-sense-and-nonsense/

[6] Redlands ESRI. 2011. ArcGIS desktop: release 10. *Environmental Systems Research Institute, CA* (2011).

[7] Bob Franklin. 2004. Talking Past Each Other: Journalists, Readers and Local Newspaper Reporting of General Election Campaigns in the UK. *Journal of Public Affairs* 4, 4 (2004), 338–346. https://doi.org/10.1002/pa.196

[8] Bob Franklin (Ed.). 2006. *Local Journalism and Local Media: Making the Local News.* Routledge. OCLC: ocm63808022.

[9] Camilla Gant and John Dimmick. 2000. Making Local News: A Holistic Analysis of Sources, Selection Criteria, and Topics. *Journalism & Mass Communication Quarterly* 77, 3 (2000), 628–638. https://doi.org/10.1177/107769900007700311

[10] Pengjie Gao, Chang Lee, and Dermot Murphy. 2019. Financing Dies in Darkness? The Impact of Newspaper Closures on Public Finance. *Journal of Financial Economics* (2019). https://doi.org/10.1016/j.jfineco.2019.06.003

[11] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H Chi. 2011. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the SIGCHI conference on human factors in computing systems.* ACM, Vancouver, Canada, 237–246.

[12] Kristy Hess. 2013. BREAKING BOUNDARIES: Recasting the "Local" Newspaper as "Geo-Social" News in a Digital Landscape. *Digital Journalism* 1, 1 (2013), 48–63. https://doi.org/10.1080/21670811.2012.714933

[13] Kristy Hess and Kathryn Bowd. 2015. Friend or Foe? Regional Newspapers and the Power of Facebook. *Media International Australia* 156, 1 (2015), 19–28. https://doi.org/10.1177/1329878X1515600104

[14] Kristy Hess and Lisa Waller. 2014. Geo-Social Journalism: Reorienting the Study of Small Commercial Newspapers in a Digital Environment. *Journalism Practice* 8, 2 (2014), 121–136. https://doi.org/10.1080/17512786.2013.859825

[15] Matthew Hindman. 2011. Less of the Same: The Lack of Local News on the Internet. FCC. , 39 pages. https://docs.fcc.gov/public/attachments/DOC-307476A1.pdf

[16] Brett Hutchins. 2004. Castells, Regional News Media and the Information Age. *Continuum* 18, 4 (2004), 577–590. https://doi.org/10.1080/1030431042000297680

[17] R Kirkpatrick. 2001. Are Community Newspapers Really Different? *Asia Pacific Media Educator* 1, 10 (2001), 16–21.

[18] Wilson Lowrey, Amanda Brozana, and Jenn B. Mackay. 2008. Toward a Measure of Community Journalism. *Mass Communication and Society* 11, 3 (2008), 275–299. https://doi.org/10.1080/15205430701668105

[19] Maxwell E. McCombs and James P. Winter. 1981. Defining Local News. *Newspaper Research Journal* 3, 1 (1981), 16–21. https://doi.org/10.1177/073953298100300103

[20] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist. 2011. Understanding the demographics of twitter users. In *Fifth international AAAI conference on weblogs and social media.*

[21] Damian Radcliffe and Christopher Ali. 2017. Local News in a Digital World: Small-Market Newspapers in the Digital Age. Tow Center for Digital Journalism. , 114 pages. https://academiccommons.columbia.edu/doi/10.7916/D8WS95VQ

[22] Meghan E. Rubado and Jay T. Jennings. 2019. Political Consequences of the Endangered Local Watchdog: Newspaper Decline and Mayoral Elections in the United States. *Urban Affairs Review* (2019). https://doi.org/10.1177/1078087419838058

[23] Lee Shaker. 2014. Dead Newspapers and Citizens' Civic Engagement. *Political Communication* 31, 1 (2014), 131–148. https://doi.org/10.1080/10584609.2012.762817

[24] Stefan Wojcik and Adam Hughes. 2019. Sizing up Twitter users. *Washington, DC: Pew Internet & American Life Project. Retrieved May* 1 (2019), 2019.