FiDO: A Community-based Web Browsing Agent and CDN for Challenged Network Environments

MORGAN VIGIL-HAYES and ELIZABETH BELDING, University of California, Santa Barbara,, USA
ELLEN ZEGURA, Georgia Institute of Technology,, USA

Homes located on tribal lands, particularly in rural areas of the United States, continue to lack access to broadband Internet and cellular connectivity [19]. Inspired by previous observations of community content similarity in tribal networks, we propose FiDO, a community-based Web browsing and content delivery system that takes advantage of user mobility, opportunistic connectivity, and collaborative filtering to provide relevant Web content to members of disconnected households via opportunistic contact with cellular base stations during a daily commute. We evaluate FiDO using trace-driven simulations with network usage data collected from a tribal-operated ISP that serves the Coeur d'Alene Indian Reservation in Western Idaho. By collecting data about household Web preferences and applying a collaborative filtering technique based on the Web usage patterns of the surrounding reservation community, we are able to opportunistically browse the Web on behalf of members of disconnected households, providing an average of 69.4 Web pages (all content from a specific URL, e.g., "http://gis.cdatribe-nsn.gov/LandBuyBack/") crawled from 73% of their top 10 most visited Web domains (e.g., "cdatribe-nsn.gov" or "cnn.com/") per day. Moreover, this content is able to be fetched and pushed to users even when the opportunistic data rate is limited to an average of only 0.99 Mbps ($\sigma = 0.24$ Mbps) and the daily opportunistic connection time is an average of 45.9 minutes ($\sigma = 2.3$ minutes). Additionally, we demonstrate a hybrid "search and browse" approach that allocates a percentage of opportunistic resources to the download of user-specified content. By dedicating only 10% of opportunistic windows of connectivity to the download of social media content, 51% of households were able to receive all of their daily expected social media content in addition to an average of 55.3 Web pages browsed on their behalf from an average of 4 different Web domains. Critically, we demonstrate the feasibility of a collaborative and community-based Web browsing model that extends access to Web content across the last mile(s) using existing infrastructure and rural patterns of mobility.

ACM Reference Format:
Morgan Vigil-Hayes, Elizabeth Belding, and Ellen Zegura. 2017. FiDO: A Community-based Web Browsing Agent and CDN for Challenged Network Environments. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 108 (September 2017), 25 pages.
DOI: http://doi.org/10.1145/3132030

1 INTRODUCTION

Approximately 53.1% of people worldwide continue to lack access to the Internet [26]. Communities located in rural and developing areas are most drastically impacted, as limited infrastructure and population scarcity limit the effectiveness of even the most robust solutions. Communities located in developed countries are not immune to this lack of Internet access. In 2015, the FCC estimated that only 15% of people living on Native American reservations in the U.S. had access to the Internet via fixed broadband *or* mobile broadband, citing reasons for lack of connectivity such as rural locale, rugged terrain, limited supporting infrastructure, and lack of

2474-9567/2017/9-ART108 $15.00
DOI: http://doi.org/10.1145/3132030

economic viability (from the perspective of major commercial service providers) [19]. Despite the fact that Internet and cellular connectivity are not usually available to their households, rates of smartphone ownership are not significantly different in Native American communities when compared to the general U.S. community [36, 43, 47].

Often, people living in these disconnected communities rely on Internet hot-spots (Wi-Fi Internet access), typically located in cafés, schools, places of work, or media centers, in order to access broadband Internet or cellular data connectivity available in more populated areas. In addition to these locations, cellular connectivity can often be found along major traffic corridors. Solutions for maximizing connectivity in communities that lack ubiquitous Internet access are proposed in bodies of work that explore delay tolerant networks (DTNs) [4, 18, 38, 41, 42] and Internet cafés [9, 10, 22]. With this previous work, users receive content from the Internet via *user-initiated encounters*, wherein a user initiates content delivery because they manually connect their device to the Internet or initiate content downloads because they recognize that they are connected to the Internet. In contrast, users can also receive content from the Internet via *opportunistic encounters*, wherein content is downloaded when devices automatically establish a connection to the Internet (i.e., associate with a cellular base station) and download content without any user involvement. While both models of connecting to the Internet are critical for largely disconnected communities, we focus on opportunistic encounters

Previous work that focuses on opportunistic connectivity emphasizes *search activity* [9, 22, 41]. With search activity, the information objective is well-defined: a user wants a specific piece of information and has explicitly set certain parameters that allow systems to search for that specific information. For instance, a user searching for a guacamole recipe might search using a keyword query (e.g., “guacamole recipe") or he might search using a specific URL (e.g., “http://www.foodnetwork.com/recipes/alton-brown/guacamole-recipe"). Chen et al. demonstrate how search activity can be translated to a delay tolerant model of networking, such that queries can be composed offline and dispatched opportunistically, collecting specific information on behalf of a user over time [9].

Browsing activity is distinctive from search activity. With browsing activity, the information objective is not well-defined. Users may begin browsing on a favorite Web page or smartphone app and then click through linked content as they encounter the content, without a goal more specific than encountering “interesting" information [47]. For example, a user browsing through their news feed on the Facebook smartphone app might encounter a link to an interesting article hosted at “cnn.com". After clicking on the link that takes them to the article, they might encounter links for other interesting content hosted by “cnn.com". At the end of this browsing session, a user may have encountered a number of articles, videos, songs, or other forms of Web content in an *ad hoc*, interest driven manner.

In contexts where Internet access is ubiquitous, research has explored various methods of pre-fetching, filtering, and recommending content based on user preferences and predictive models of user behavior. Web browsing agents [2, 10, 32] and Web content recommender systems [40] are among such technologies. As the Web becomes increasingly personalized, recommender systems are commonly integrated into individual Web sites and services so users can immediately access relevant content rather than spending time browsing for it. While these approaches provide a solution for computer-assisted Web browsing, they are based on individual browsing patterns and information interests. In previous work, Web browsing is performed by an individual for their own *ad hoc* informational interests. Previous work also assumes seamless access to the Web [2, 10, 32, 40]. In this work, we extend the concept of Web browsing agency to account for a different profile of information needs. First, we identify the need for agents that browse on behalf of a group of individuals (i.e., members of a household). Second, we recognize that without a historic record of the content that comprises a group’s Web browsing activity (because there is no access to the Web at home), we can utilize collaborative recommender system techniques that leverage the browsing patterns of similar entities in order to predict the browsing behaviors of members of disconnected households. In this work, we also extend traditional Web browsing agents so that they operate in a manner that enables users to opportunistically take advantage of the recommendations made on behalf of members of their household. We accomplish this by proactively fetching and caching recommended content so that it is stored at community-operated cellular base stations. Ultimately, this recommend-fetch-store process maximizes the value of regularly encountered opportunistic cellular data connections and we fill a gap in the research that allows us to make browsing more pervasive and accessible to disconnected homes.

Given the need for greater utility of opportunistic Internet access in communities where most homes lack access to the Internet, the tools provided by recommender systems, and observations from our previous work that suggests members of the same community have similar content interests [54, 56] we ask the following research questions: (i) *How much of a household's Web browsing needs can be met opportunistically?* and (ii) *To what degree can the members of a disconnected household rely on their surrounding community to identify relevant and interesting content on their behalf?*

In order to discern the extent to which community content preferences can guide content prioritization schemes necessary to deliver Web content opportunistically, we analyze Web traces collected from the Red Spectrum Communications network, a tribal operated Internet service provider (ISP) that operates within six towns in the Coeur d'Alene Indian Reservation. The Red Spectrum network connects 530 households and 42 municipality buildings to the Internet. For this study, we have collected data from 82.9 million HTTP/S transactions that represent all network Web activity between January 17 and February 28, 2017.

Motivated by our prior observations of locality of interest within the tribal networks [54, 56] and the combined potential of opportunistic networking and recommender systems, our paper makes the following contributions:

- We analyze Web usage in a tribal-operated network that connects 530 households to the Internet. We find that while household-level content similarity is low, town-level content similarity is high enough to provide a basis for content recommendations.
- We propose a community content delivery network (CDN) that operates opportunistically in a challenged network environment and functions as an agent that fetches content on behalf of an entire household. We structure a regional CDN node to proactively push content to the devices of users from disconnected homes. We populate the CDN with content cached from community Web browsing sessions that take place in areas of the community that do have Internet access (e.g., schools, libraries, and homes).
- We evaluate the performance of our proposed system using trace-driven simulations. Because the performance of CDNs are highly dependent on the specifics of traffic access, our unique access to traffic traces provides critical realism in our evaluation. While user mobility models have been established for metropolitan contexts [27], these models are not well-suited to the realities of mobility through rural and rugged terrain. In order to better simulate user mobility in under-studied, rural communities, we create a mobility model based on census and transportation data.

While we focus our work on tribal reservations due to our current partnerships, our work is more broadly applicable to rural communities in general.

2 RELATED WORK

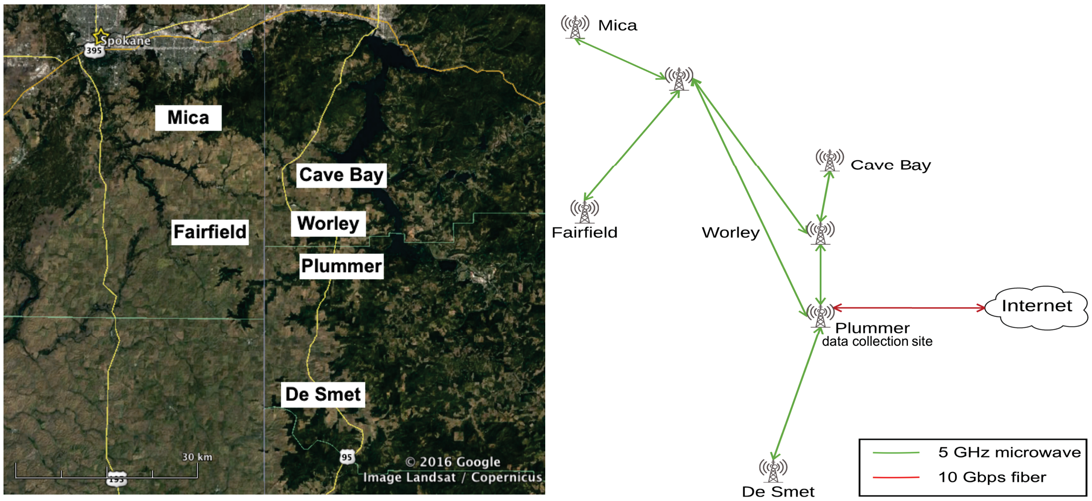
Since the inception of delay tolerant networking (DTN) [18], DTN research has focused on developing applications for operation in challenged network environments [4, 38], DTN architectures [4, 38], and routing solutions for DTN architectures [28, 59]. Work that proposes architectures for Web search over DTNs is of particular relevance. Ott et al. propose a bundling solution that enables HTTP transactions to take place over intermittent connectivity [38]. Balasubramanian et al. propose an architecture that utilizes urban buses as message store-and-forward nodes for Web search transactions [4]. Specifically relevant to this study is research that investigates DTN operation in rural and developing contexts [9, 41, 46], wherein intermittent connectivity is utilized to fetch and post information on behalf of users who are typically disconnected. Pentland et al. challenge the myth that connectivity must be real-time in order to be relevant to meet rural community needs, and demonstrate how the combination of asynchronous connectivity and wireless access points can sufficiently serve developing communities in rural India [41]. Similarly, Seth et al. propose the usage of mechanical backhauls in lieu of expensive wireless connectivity solutions (e.g., VSAT) to connect rural information kiosks to the Internet. The system we propose is distinct from this prior work in that our work focuses on proactively collecting and delivering content on behalf of users without requiring explicit user-prompted searches or requests. We also focus on delivery of content directly to rural households, rather than to specified kiosks.

In the past decade, information and communication technology for development (ICTD) research agendas have identified the sustainability, relevance, and affordability of local, community-based networks [24, 29, 30, 45, 60]. In particular, studies of network usage in rural communities have demonstrated a locality of interest with respect to the content that is downloaded [29, 30, 45] and interacted with online [29, 45, 56]. Previous work has demonstrated how this locality of interest can be exploited by local, community-based networking architectures to provide a myriad of services that maximize network resources in networks that may be prone to disconnection and gateway congestion. Some examples of community networks include small-scale, community-operated cellular networks [23, 60] and community-based content caches for videos and images uploaded by community members [30, 45]. The system described in this study is inspired by these previous observations and community-based architectures. Specifically, we take advantage of the rise of community-based wireless Internet Service Providers (ISPs) to propose a comprehensive content delivery system that targets members of disconnected homes in partially connected communities. Community-based ISPs present a unique opportunity to re-imagine information systems. Often, community-based ISPs have been established in response to unmet community information needs, including general lack of broadband Internet availability and lack of affordable service options [16]. By working alongside these community-oriented efforts, we are afforded the unique opportunity to re-imagine the utilization of ICT resources and maximize their value by making community information needs and usage patterns a focal point of system design.

Also related to our work are recommender systems, especially those that rely on collaborative filtering [44] and demographic-based [39] recommendations. Our work is particularly inspired by personal Web browsing agents and automated Web search [3, 11, 32, 40]. These browsing agents rely on content-based recommendation techniques and user behavior in order to initiate relevant Web searches on behalf of users while they simultaneously browse the Web. Similarly, our system initiates Web searches on behalf of the user, though these searches take place asynchronously and opportunistically, rather than parallel to the user's Web browsing activity.

3 BACKGROUND

Only 15% of Native Americans residing on tribal lands have access to the Internet [19]. Most of the challenges for broadband accessibility on tribal lands stem from a centuries-long history of conflict between the U.S. government and Indigenous Americans, which resulted in the forced relocation of tribes to geographically remote areas often marked by rugged terrain. As tribes in the U.S. seek to cultivate broadband infrastructure within their communities, they must contend with national telecommunications policy, coordinate with neighboring deployment strategies and regulatory bodies, and find sustainable revenue sources. Issues of resource allocation and use are complicated by their status as sovereign nations. While tribal lands are far from experiencing ubiquitous Internet connectivity, efforts initiated by tribal communities and the Federal Communications Commission (FCC) have created a mosaic of connectivity, where connectivity is often present to some degree at the hub of tribal communities (typically in tribal offices, learning centers, and health clinics) and along the major transportation corridors that may pass by or through reservations; however, connectivity typically does not penetrate homes in the more rural parts of reservations [19].



(a)

(b)

Fig. 1. Geographic map of (a) communities connected by the Red Spectrum network and (b) a topographical map of the communities connected to the Internet over 5 GHz microwave links. All communities are part of the Coeur d'Alene Indian Reservation except Fairfield and Mica.

Table 1. Demographic and observational statistics for the towns connected by the Red Spectrum network that we observe in our study. We note that we cannot differentiate between IP addresses based in Mica and Fairfield; thus, we remove traces associated with these IP subranges for in-town traffic similarity analysis in Section 4.1.

Town	Pop. size	Observed wireless subscribers	On tribal land?
Plummer, ID	1,026	161	Yes
Fairfield, WA	606	104	No
Mica, WA	563	NA	No
Worley, ID	254	117	Yes
Rockford Bay, ID[1]	184	24	Yes
De Smet, ID	175	23	Yes

3.1 Network overview

Our study is based on data collected from a network that serves tribal communities operating in predominantly rural and resource-challenged environments. Taken as a whole, Native American reservations represent a community that is distinct from the general U.S. population. This has implications for the potential overlap in content preferences that may be present within these culturally similar communities [56].

The Red Spectrum network has so far deployed infrastructure that provides Internet access to the approximately 7,000 homes located throughout the six communities that comprise the Coeur d'Alene Indian Reservation in western Idaho. Since its inception, the Red Spectrum network has provided Internet to 1,011 subscribers. We observed a total of 530 active subscribers during our collection period, 43 of which are municipal subscribers (e.g., schools, tribal offices, libraries, and health clinics). We map the towns that comprise the Red Spectrum network in Figure 1a.

Table 2. An overview of the data sets collected from the Red Spectrum network.

Subnetwork	Observed wireless subscribers	Total traffic volume (TB)	% volume Web traffic	# Web transactions (millions)
Plummer	161	2.8	98.3	9.6
Worley	117	3.6	97.4	12.4
Cavebay	24	0.3	95.8	1.4
De Smet	23	0.2	94.8	1.3
Mica/Fairfield	104	2.1	97.0	9.4
All wired	101	14.8	95.5	48.8
Total	530	23.9	95.7	82.9

In the Red Spectrum network, the gateway delivers 10 Gbps over a fiber link that terminates at the Red Spectrum headquarters located in Plummer, ID. Connectivity is extended to residents either through a wireless backbone comprised of 5 GHz microwave links or via fiber links deployed directly to the subscriber. We provide a map of the network topology in Figure 1b.

3.2 Data collection

Our point of collection is located at the Internet gateway in Plummer (see Figure 1b). We collect data by attaching a traffic monitoring server to the switch that bridges the gateway and the network. A mirror port is configured to capture all packets traversing the network. We use the Bro Network Security Monitor to collect flow-level statistics for all Web traffic that traverses the gateway link [49]. We observed 23.9 TB of network traffic, 246 billion packets, and 82.9 million HTTP/S transactions in the Red Spectrum network during our collection period. We report general statistics associated with each of the communities that comprise the Red Spectrum network in Table 2. All MAC and IP addresses are anonymized using TraceAnon [50]. For household statistics at the aggregate level, we refer to all household usage in the network (which we discern by removing usage data generated by municipal subscribers). It is critical to note that all data collected reflects usage by homes that are connected to the Internet.

4 CONTENT PREFERENCES

In order to begin characterizing the feasibility of leveraging opportunistic connectivity and community similarities to browse the Web on behalf of members of disconnected households, we analyze the content preferences of households and compare them to the aggregate preferences of their surrounding community.

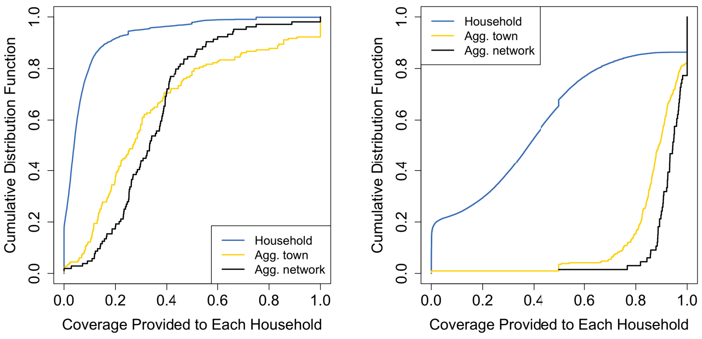
4.1 Web preferences

To understand how well community Web usage represents household Web usage, we compare the similarity of Web preferences among communities of various scopes: global, town, and household. *Global* preferences are determined using Web traffic from the entire Red Spectrum network; *town* preferences are determined using Web traffic generated by the six individual townships located within the Coeur d'Alene Indian Reservation; and *household* preferences are determined using Web traffic generated by each household.

We begin our analysis of community preference similarities by calculating content coverage using

$$C(A,B) = \frac{A \bigcap B}{A} \tag{1}$$

[1] Served by the Cavebay tower.



(a) (b)

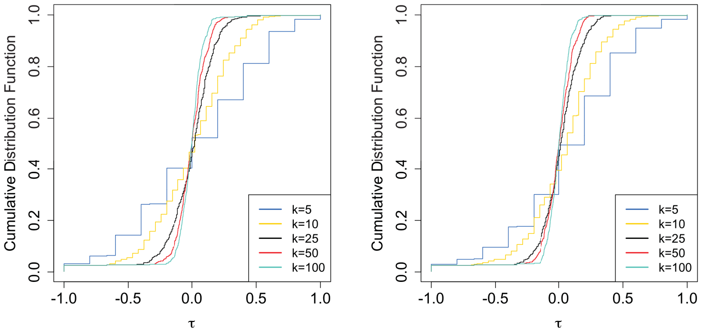
Fig. 2. Cumulative distributions associated with the (a) file coverage provided by different scopes of community and (b) domain coverage provided by different scopes of community in the Red Spectrum network.

where A represents the content accessed by a particular household and B represents the content accessed by some other group (i.e., another household, the corresponding town, or the global network). When we calculate coverage provided for a household at the town or global level, we remove that household from the aggregate coverage at the town level or the global network level. Coverage ranges from 0 (where B has not accessed any of the content accessed by A) to 1 (where B has accessed all content accessed by A).

We plot coverage with respect to downloaded files in Figure 2a, which represents the cumulative distribution of file coverage provided to each household by other individual households in the same town (“Household"), by the aggregate town community (“Agg. town"), and by the global community (“Agg. network"). We find that while the mean file coverage provided by other individual households in the town is only 0.07 ($\sigma = 0.12$), the mean coverage provided by the aggregate town community is 0.35 ($\sigma = 0.28$). Based on a two-sample Kolmogorov-Smirnov test, we observe that the file coverage provided by the global community is not significantly greater than that provided by the town community ($p < 0.001$). This leads us to believe that curating a community content delivery system based on global Web usage would not significantly outperform the same system based on town Web usage with respect to the files stored for delivery.

In addition to file coverage, we seek to characterize the community’s ability to provide Web content for disconnected households by examining the *domain* coverage provided at different scopes. Instead of measuring coverage on specific file content, we measure it with respect to the Web domains visited by individual homes and the surrounding community. Since we only have flow-level information about household Web usage, domain is a proxy for content interest; even if the accessed files are different, files from the same domain serve as a heuristic for recommendation. Figure 2b plots the cumulative distributions associated with domain coverage provided to each household by other households in the same town (“Household"), by the household’s town in aggregate (“Agg. town"), and by the entire Red Spectrum network (“Agg. network"). While individual households do not provide significant coverage to each other (mean coverage is 0.4 ($\sigma = 0.32$)), communities do provide significant domain coverage (mean coverage at the town level is 0.87 ($\sigma = 0.14$) and 0.93 ($\sigma = 0.07$) at the global level). Thus, relying on the aggregate community to source a household’s Web content interests, at the domain level, is quite plausible.

As we look to filter content based on community popularity, we need to identify the scope of community that ranks content most similarly to individual households. Preference similarity at the level of files is very fine-grained and may be so precise as to be prohibitive for filtering a large number of files with similar community



(a) (b)

Fig. 3. Cumulative distribution associated with Kendall's τ similarity between top k household domains and (a) top k town domains and (b) top k global domains in the Red Spectrum network.

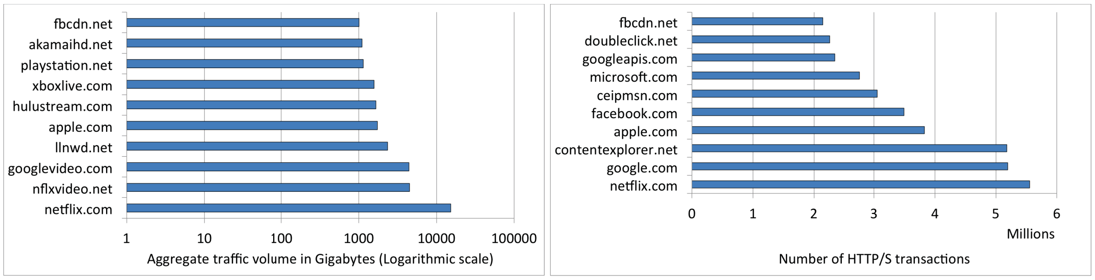
rankings. We address this by examining similarity at the level of Web domains, which we rank according to the number of files downloaded by each household from the domain during our observation period. Using this method, domains that are associated with a larger number of files downloaded by a household are ranked higher than those associated with a smaller number of files downloaded by a household. We then compare domain ranks for the top k domains using the Kendall τ rank correlation coefficient [17], which is calculated by:

$$\tau = \frac{\#\ concordant\ pairs - \#\ disconcordant\ pairs}{k(k-1)/2} \tag{2}$$

where k is the number of items ranked in the list and τ ranges from -1 (completely different rankings) to 1 (identical rankings). Concordant pairs represent two domains with the same relative rank. For example, if domain x is ranked higher than domain y in Lists 1 and 2, then domain x and domain y are considered a concordant pair; otherwise, they are considered a disconcordant pair. We use Kendall's τ rank correlation instead of other rank similarity metrics (such as Spearman's ρ) because it provides a more direct interpretation of similarity based on the presence of concordant pairs and it takes tied ranks into account.

We compare the ranking of the top k domains for each household with the aggregate top k domains for the entire Red Spectrum network (global), with the aggregate top k domains for the corresponding town, and with other households on the Red Spectrum network for $k = \{5, 10, 25, 50, 100\}$. Figure 3 plots the cumulative distribution of the Kendall τ correlation for these comparisons. For each scope of comparison, the mean correlation decreases as k increases. However, we find that the greatest correlation occurs between the top k household domain ranks and the aggregate domain ranking of the household's corresponding town. For this comparison, the mean Kendall τ correlation is 0.04 for $k = 5$. A two-sample Kolmogorov-Smirnov test between the Kendall τ distributions associated with the town level comparison and the global level comparison reveal a significant difference at the $p \leq 0.05$ level between the correlation at these different scopes for $k = 5$ but not for other values of k. Therefore, if a user has only a limited amount of time to opportunistically download content on behalf of their household, their content interests would be better served by downloading content that has been ordered according to the aggregate rank provided by their town community, rather than the aggregate rank provided by the entire network. Conversely, this demonstrates that for lengthier opportunities, users may be equally served by aggregate rankings established at either a town or global level, particularly when considered in conjunction with our findings of greater domain coverage at the global level.

4.2 Social media preferences



(a) (b)

Fig. 4. Top ten Web domains based on (a) aggregate traffic volume and (b) total number of HTTP/S transactions over the observation period.

Online social networks (OSNs) play a critical role in the geopolitical landscapes of challenged network environments, particularly when a lack of Internet access results from political marginalization and aggravates cultural marginalization. This is particularly true for Indian Country; survey and interview studies have demonstrated that OSNs are the most popular means of everyday communication in tribal communities [7, 20, 35]. Moreover, OSNs play a potentially powerful role in Indigenous cultural preservation, revitalization, and political empowerment [7, 12, 20, 31, 35, 58]. Indeed, our own previous work highlights the importance of OSN content in Indian Country as a platform by which Native Americans strengthen community bonds [55, 56] and form communities via identity-driven content [57].

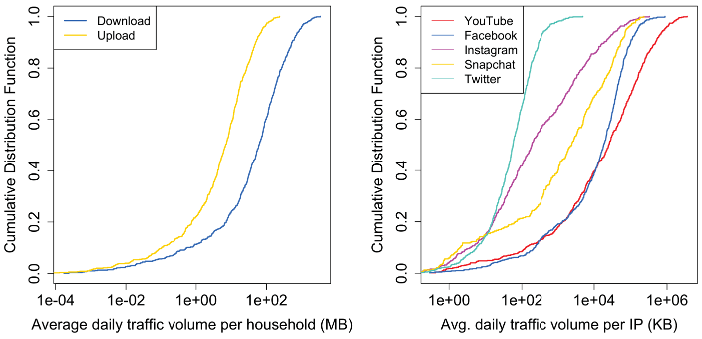
When examining the Web traffic profile of the Red Spectrum network in Figures 4, we find that Facebook (“facebook.com" and “fbcdn.com") is the most accessed social media site as well as the social media site associated with the greatest volume of traffic over our observation period. Indeed, if HTTP/S requests generated by “facebook.com" and “fbcdn.com"[2] were aggregated to count as a single domain, Facebook Web transactions would outnumber Netflix transactions by over a hundred thousand during our collection period. Understanding social media usage is critical to evaluating the feasibility of whether opportunistic cellular connectivity accessed while commuting would be sufficient to deliver the social media demands of all members of a household. Moreover, we differentiate social media traffic from other traffic as it has the potential to provide insight into social connectivity, which can be used for social-based content recommendations assuming proper information access [54]. While we do not leverage social connectivity for content recommendations in this work, we do understand that supporting social media content is a crucial first step towards the integration of more sophisticated approaches for browsing recommended Web content. To this end, we perform a deeper analysis on social media platform usage in the Red Spectrum network by comparing the usage of the top 5 most accessed social media platforms, including the penetration rates (reported in Table 3) and daily traffic volume (plotted in Figure 5).

We find that the majority of households in the Red Spectrum network access Facebook, YouTube, Twitter, or Instagram during our observation period, while over one-third of users access Snapchat in the same time frame. Given the prevalence of these platforms, we measure the associated traffic volume associated with each IP address for all social media platforms in Figure 5a. We find that the average household downloads a median of 61.1 MB (σ = 373.2 MB) of social media content and uploads a median of 6.8 MB (σ = 30.6 MB) to social media platforms on a daily basis. When we divide household social media traffic based on platform, we find that the

[2] Domain associated content stored on Facebook’s content delivery servers.

Table 3. Percentage of IP addresses that access the top 5 social media platforms accessed by the Red Spectrum network.

Social media platform	% of households that have accessed during observation period
Facebook	76.7
YouTube	74.1
Twitter	74.1
Instagram	64.8
Snapchat	37.2



(a)

(b)

Fig. 5. Cumulative distribution of the (a) average daily traffic volume for all social media and (b) for the top five most accessed social media platforms accessed by users in the Red Spectrum network.

greatest traffic volume is associated with YouTube and Facebook followed by Snapchat, Instagram, and Twitter (see Figure 5b).

5 FIDO OPERATION

As previously stated, while many households on reservations lack both Internet and cellular connectivity, cellular coverage is often available in select locations, such as along major traffic corridors or in municipal areas. Our goal is to push relevant content to users through their cellphones as they pass through these areas of coverage, enabling individuals to collect Web content on behalf of themselves and other members of their household in a way that *(i) prioritizes the most relevant content* and *(ii) initiates collection of Web content without requiring explicit user interaction during moments of connectivity.*

To help members of disconnected households in rural areas of reservations take advantage of the opportunistic connectivity that they encounter as they mobilize throughout their day, we propose FiDO (**Fi**les **D**elivered **O**pportunistically), a community-based content delivery network that pushes files downloaded by members of the surrounding community to mobile users. In the FiDO architecture, local CDN nodes are placed throughout the reservation at community-run cellular base stations and wireless ISP towers (if they exist). These nodes coordinate with a regional content store (which may be placed, for instance, at the tribal headquarters of each reservation or some other municipal building in the rural community) that pushes new content to the CDN nodes and stores a copy of content requested at each CDN node at regular intervals. We illustrate the usage scenario in Figure 6. Here, we show a mobile user associated with a disconnected household. As the user travels away from home and throughout her reservation (e.g., going to work or school), she encounters areas of cellular

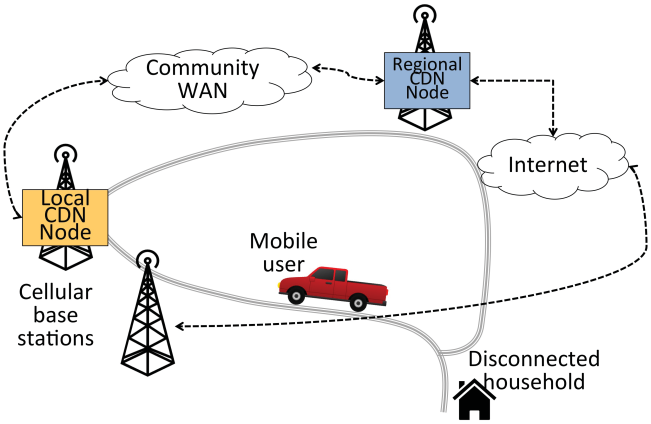


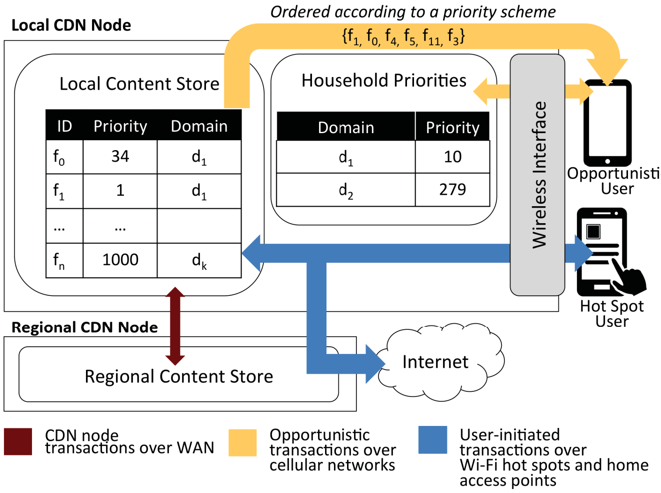
Fig. 6. An example of FiDO's operation, where a member of a disconnected household opportunistically collects relevant content from CDN nodes located on cellular base stations along their commute.

connectivity placed along major traffic corridors. If these cellular base stations are part of a community network and are associated with a local CDN node, the node will push content to the user; otherwise, if the base stations are commercial, users connect to their regional CDN node over the Internet.

In Figure 7, we illustrate the FiDO data flow process. The main components include local CDN nodes, user devices, and regional CDN nodes. Regional CDN nodes contain a content store of all files that have been downloaded in their region over a period of time. A control protocol based at regional CDN nodes synchronizes content on all local CDN nodes at regular intervals so that each CDN node has a copy of all the files that have been downloaded regionally within a 24 hour period. Local CDN nodes include a content store and, depending on the prioritization scheme used, a database containing household preferences. As users throughout the network browse and search the Web (user-initiated transactions), a copy of the files they access are stored at both the regional and local CDN nodes in the network. Conversely, if a user is opportunistically connected to a local CDN node, content is pushed to the user's device according to a prioritization scheme (opportunistic transaction). Depending on the prioritization scheme used, a local CDN node may also request a list of Web preferences associated with the user's household (see Section 6.2 for a description of prioritization schemes). These preferences are then used to tailor the prioritization scheme to best accommodate the content needs of the household. To prevent the same file from being transmitted to a mobile user multiple times throughout the day, as the user moves and changes her point of attachment, user devices transmit a list of file identifiers already received to the CDN node, so that the prioritization scheme always schedules new content. We note that this type of system requires an application running on mobile user devices that would enable households to share preferences and allow users to share fetched content. These modifications are discussed in Section 7.

6 EVALUATION

In this section, we use trace driven simulations to evaluate how well FiDO provides households with relevant content for the day. Our first goal is to quantify the potential that opportunistic cellular connections have for


Ordered according to a priority scheme
$\{f_1, f_0, f_4, f_5, f_{11}, f_3\}$
Local CDN Node
Local Content Store
Household Priorities
Wireless Interface
Opportunistic User
Hot Spot User
Regional CDN Node
Regional Content Store
Internet
CDN node transactions over WAN
Opportunistic transactions over cellular networks
User-initiated transactions over Wi-Fi hot spots and home access points

Local Content Store

ID	Priority	Domain
f_0	34	d_1
f_1	1	d_1
...	...	
f_n	1000	d_k

Household Priorities

Domain	Priority
d_1	10
d_2	279

Fig. 7. FiDO data flow diagram. Arrows represent the flow of content. Content is browsed by users connected to the Internet at home or at WiFi hot spots. FiDO fetches and stores content (which has been filtered using the browsing patterns of the surrounding community) on behalf of disconnected households. When a user from a disconnected household connects opportunistically, FiDO pushes content to the user's device according to a prioritization scheme.

delivering a household's daily Web browsing needs. Our second goal is to characterize how well browsing patterns of the surrounding community can inform the browsing interests of disconnected households.

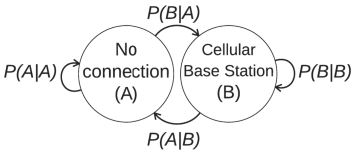
6.1 Simulation overview

Our evaluation of FiDO relies on a trace-driven approach that allows us to measure FiDO's ability to meet household content needs using actual usage data. Ultimately, we simulate the scenario outlined in Figure 6, where a user collecting content on behalf of their household encounters areas of cellular connectivity as part of their normal commute and opportunistically downloads Web files on behalf of members in their household.

To simulate user mobility through alternating areas of cellular coverage, we rely on a two-state Markov model (shown in Figure 8a) that transitions between states of coverage, where state A represents a lack of Internet coverage and state B represents mobile broadband connectivity via a cellular base station.

Since FiDO has been designed specifically to deliver content in rural communities with limited access, we simulate user commutes through rural areas using rural transportation statistics. Specifically, studies by the U.S. Department of Transportation have found that for rural residents, the average number of driving minutes[3] per day is 55.87 minutes [14]. We assume that this average number of driving minutes occurs at the average rural speed limit of 75 miles per hour [8] and we assume that the one way driving time (i.e., the time to commute to work) is half the total daily driving minutes, or 27.94 minutes. In order to simulate comparable commute times, we select each user's daily commute time from a normal distribution with a mean of 27.94 minutes and a standard deviation of 5 minutes in our baseline models. We restrict a user's driving times to occur between typical commute hours (7 to 9 a.m. and 5 to 7 p.m.) [14]. All users begin in a state of disconnection and transition to a state of connectivity depending on the time of day as well as the amount of time, t, they have already traveled in the simulation period. Figure 8b presents the transition probabilities based on the time of day. The one-way commute time, c_w, is modeled from a normal distribution with a mean of 27.94 minutes and a standard deviation

[3] *Driving minutes* refers to minutes spent driving on the road.



(a)

Home or Business Hours (7 p.m. to 7 a.m. and 9 a.m. to 5 p.m.)	Commuting Hours (7 to 9 a.m. and 5 to 7 p.m)
$P(B \mid A) = 0$ $P(A \mid B) = 1$ $P(B \mid B) = 1 - P(B \mid A)$ $P(A \mid A) = 1 - P(A \mid B)$	$P(B \mid A) = \begin{cases} 1, & t < c_w \\ 0, & t \geq c_w \end{cases}$ $P(A \mid B) = \begin{cases} 0, & t < c_w \\ 1, & t \geq c_w \end{cases}$ $P(B \mid B) = 1 - P(B \mid A)$ $P(A \mid A) = 1 - P(A \mid B)$

(b)

Fig. 8. (a) Connectivity state machine used in simulation. (b) State machine transition probabilities based on the time of day.

of 5 minutes. We note that t is reset to $t = 0$ when a user reaches a disconnected state during Home or Business hours.

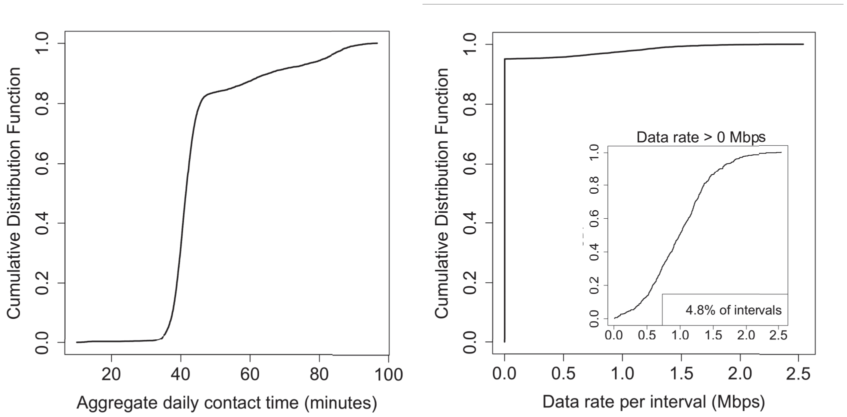
Our simulations are run over seven representative days of data collected between February 1 and February 7, 2017. The simulation is run in one minute intervals, meaning that connectivity state and data rate is evaluated for every simulated minute. In order to evaluate FiDO's performance for disconnected households, we randomly select households from three of the communities in the Red Spectrum network to emulate the desired content of disconnected households in our trace-based simulation. For each run through the simulation we select 10 households and we run the simulation five times with random seeds for each community. For selected households, we use the traces of their Web usage to function as a ground truth with respect to the actual files they expect to receive and the times they expect to receive them. On average, each household requests 39.2 files ($\sigma = 472.4$) daily. Our evaluation specifically simulates a user from each household opportunistically collecting content on behalf of the household; thus, we simulate mobile users to correspond to each of the selected households.

In order to simulate access restrictions associated with specific Web files, users can only receive a Web file if they have actually received it in the actual traces of use. Our simulations assume that members of a household can entrust their access credentials to the user who is connecting to FiDO on their behalf. The significance and complications of these assumptions are discussed in Section 7.

6.2 Prioritization schemes

At the most general level, the system outlined in Figure 6 selects content from content stores located within local CDN nodes in order to opportunistically provide content to a user on behalf of her household. Inevitably, the content stores contain more content than can be transferred during opportunistic encounters and not all content stored is relevant to every household. Therefore, we propose and evaluate several prioritization schemes that are used to select and prioritize content that is to be transferred during opportunistic connections.

6.2.1 Naive scheme. The *naive prioritization scheme* (denoted as “Naive" in evaluation graphs) relies on collaborative filtering for content selection and does not take into account the preferences of individual households.



(a)

(b)

Fig. 9. Distribution of the (a) daily contact time users have with a cellular base station and over the course of a simulation run and (b) the average data rate a user is connected by for each minute interval in the simulation. The inset in (b) graphs the distribution of the 4.8% of intervals where the user is connected to a cellular base station.

In this way, the naive scheme represents a system where the system infrastructure operates independently of the users who are opportunistically connecting. Files are prioritized based on the number of times community members download the file within a moving time window of 24 hours.

6.2.2 User preference scheme. The *user preference prioritization scheme* (“User Pref.") uses the domain preferences of household members to impose an additional prioritization that operates on top of the collaborative filtering used in the naive scheme. When users connect to a local CDN node, they provide a domain preference list. In practice, this ranked list could be generated explicitly by users in a household or implicitly based on usage. We simulate household preferences as a list that ranks domains according to the historic number of files a household downloads from the domain (i.e., a domain from which 1,000 files are downloaded ranks higher than a domain from which 100 files are downloaded). While an opportunistic connection to a CDN node exists, the domain preference prioritization scheme cycles through each domain from highest ranked to lowest ranked. As the prioritization scheme comes to each domain, it pushes the files downloaded from that domain based on the number of times the community has downloaded the file.

6.2.3 Push-pull scheme. The *push-pull prioritization scheme* (“Push/Pull") creates two prioritization queues. When a user connects to a local CDN node, her device makes requests for specific files on behalf of her household. If the file is already stored at the local CDN node, it is pushed immediately to that user and removed from the user’s request queue. Otherwise, a pull request is made and the file is downloaded from the Web and made available to every local CDN node in the region within the next 10 minutes (to simulate synchronization latency). In the meantime, the user adds the file identifier to the request queue, which is ordered on a first-come, first-served basis. At each opportunistic connection, the request queue is serviced first. When the request queue is empty, the push-pull prioritization scheme operates identically to the user preference scheme. We note that the push-pull scheme serves to demonstrate an ideal scenario, where a household is able to engage in *ad hoc* Web browsing via opportunistic connectivity accessed by its mobile user. As such, the push/pull scheme is the main mechanism by which we evaluate our first research objective: *How much of a household’s Web browsing needs can be met opportunistically?*

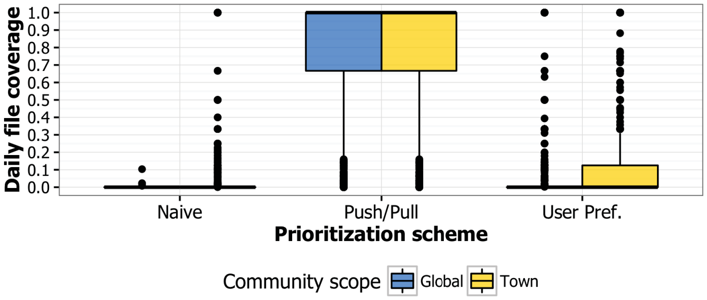


Fig. 10. Boxplot of distributions associated with daily file coverage provided by each of the prioritization schemes assuming an average total daily commute time of 55.87 minutes traveled at 75 miles per hour. We graph the coverage provided when using recommendations by the household's town in yellow and recommendations by the global network community in blue.

6.3 Filtering on cached files

Our first set of experiments evaluates FiDO's ability to predict specific Web files that members of a household would access throughout their day given only opportunistic access to Internet connectivity. In this set of experiments, FiDO stores Web files that have been downloaded by community members and pushes them opportunistically to a mobile user according to the prioritization schemes detailed in Section 6.2. A simulation run consists of 10 randomly selected users from a single town, an average commute time, and a day's worth of traffic traces. We run all simulation configurations for traces collected from the towns of Worley, Plummer, and Mica/Fairfield. The results reported are based on a total of 121 unique households, where 10 from the same town are withheld from traces used to generate the simulation results for each run. We evaluate performance of our proposed system using notions of file coverage (discussed in Section 4). Using the traces of actual household usage, we are able to compare what households operating as "offline households" would receive if they did have Internet available in their home to what they would receive using FiDO. We measure coverage provided at the end of the day as well as coverage provided by the end of the commute back home from work.

In Figures 9a and 9b, we graph the distribution of the contact time users have with a cellular base station and the distribution of the data rate available to the user in each interval over the course of a single simulation run (one day). Based on our simulation environment, users are in contact with a cellular base station for an average of 45.9 minutes ($\sigma = 2.3$ minutes) a day. For the minutes that a user is in contact with a cellular base station, they receive content at an average rate of 0.99 Mbps ($\sigma = 0.24$ Mbps).

We plot the distributions of file coverage achieved by each prioritization scheme in Figure 10. The average file coverage provided by the schemes based on collaborative filtering ("Naive" and "User Prefs.") is very low–only an average of 0.15 ($\sigma = 0.3$) for the user preference scheme and 0.04 for the naive preference scheme ($\sigma = 0.16$). This is not very surprising, as the filtering occurs over specific files that comprise a single Web page. As Web pages are increasingly dynamic and individualized, it is unlikely that a visit to the same Web page would yield the exact same files for two different individuals. Most importantly, we find that the push/pull scheme, which essentially functions as an oracle scheme (i.e., the optimal approach), provides an average file coverage of 0.80 ($\sigma = 0.36$). This means that even if the user is relying exclusively on opportunistic cellular connectivity to access the Internet (as modeled by our simulation), she will be able to collect all cacheable content her household would expect to receive during the day if they were connected to the Internet.

6.4 Filtering on crawled domains

In our analysis of Web preference similarity between households and their surrounding community, we found that while the aggregate file coverage for households averages at 0.35 with high variance ($\sigma = 0.28$), the average domain coverage provided by the surrounding community is quite high with little variance (mean coverage at the aggregate town level is 0.87 and mean coverage at the aggregate network level is 0.93). Our experiments in Section 6.3 reveal that collaborative filtering, even when directed by historic domain preferences of household users, is only able to provide a small percentage of a household's daily content interests. In order to provide greater coverage to household content interests, we evaluate FiDO using a "browsing model", wherein Web pages from the most broadly accessed Web domains are crawled and cached then pushed opportunistically to users according to the various prioritization schemes outlined in Section 6.2. Here, a *Web page* represents a collection of Web files that are rendered together by a browser to create a multimedia and interactive end-user experience. For this set of experiments we rely on traces collected between February 1 and 7, 2017 to identify the most broadly accessed Web domains as they would be filtered by each prioritization scheme in one minute intervals. Instead of caching specific Web file objects and prioritizing the order in which they are pushed to the user, we simulate crawling Web domains and caching entire Web pages that are then pushed to the user based on how each prioritization scheme filters Web domains. We use observations from several large-scale studies of the graphical structure of the Web to inform our simulation models [6, 13, 34]. Based on observations by Broder et al. and Clauset et al., we assume that the out-degree associated with each Web page follows a power-law distribution, where most pages link to only a few other pages and a few pages link to many other pages [6, 13]. In a more recent study of Web graph structure, Muesel et al. observe that the average out-degree for a Web page is 36.7 and the tail of the distribution decays at an exponent rate of 2.77 [34]. We simulate our Web crawl by modeling the number of links from the homepage of a given Web domain from the power-law distribution observed by these previous studies of the Web structure. We then model the size of each Web page to which the homepage links based on models observed in archived Web measurements, wherein the average Web page had a size of 2.35 MB during the first week of February 2017 and follows a Pareto distribution (we model with a shape where $\alpha = 2$) [25]. We note that we only simulate a crawl with a depth of 1, meaning we only simulate the download of pages directly linked to the homepage associated with a domain. We believe this approach models an approximation of Web structure that is accurate enough to allow us to measure the feasibility of leveraging opportunistic connectivity using community-based collaborative filtering. Metrics used to evaluate the performance of FiDO as a browsing agent include domain coverage (see Section 4), the number of different Web domains, the total number of Web pages, and the average domain rank of pages pushed to the user over the run of a simulation.

In Figure 11, we graph the distributions of the daily domain coverage provided by the naive, user preference, and push/pull prioritization schemes. We calculate daily domain coverage based on the different domains that a household accesses each day. Our simulations show that the push/pull approach is the optimal approach with respect to responsiveness to the daily changes in household Web domain interests. In general, prioritization schemes that filter based on aggregate network usage ("Global") outperform approaches that filter based on aggregate town usage ("Town") by a factor of 1.8. The mean daily domain coverage provided by the push/pull approach is 0.34 ($\sigma = 0.33$) with no significant difference between the distributions that filter over "Global" and "Town" community usage. The distribution of daily domain coverage values for the naive ($\mu = 0.11$; $\sigma = 0.19$) and user preference ($\mu = 0.12$; $\sigma = 0.22$) schemes are not significantly different at the $p < 0.01$ level of significance according to a two-sample Kolmogorov-Smirnov test. The domain coverage values we observe are quite low compared to what we observe in Section 4. The reason for this is that each of the prioritization schemes operates by downloading all of the crawled and cached Web pages associated with each domain as the domain is prioritized by the scheme. Ultimately, this limits the overall number of domains with Web pages to be opportunistically downloaded. In order to account for this, we introduce a round robin scheduling approach into the user preference

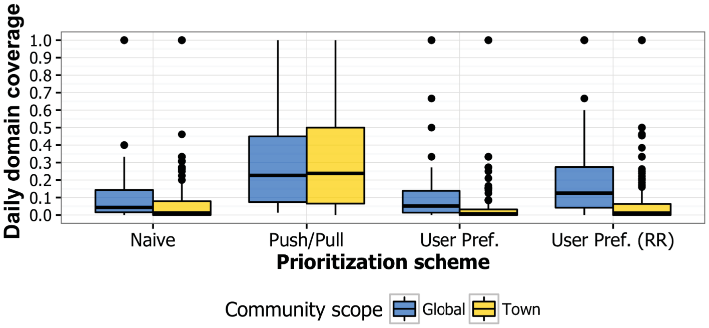


Fig. 11. Distribution of the daily coverage of Web domains expected by household members at the end of each day of the simulation. The box represents the IQR (interquartile range), the whiskers represent ±1.5× the IQR, the bold line represents the median of the distribution, and the dots represent outliers.

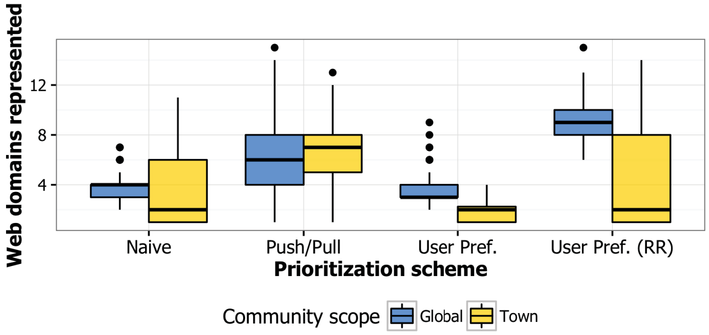


Fig. 12. Distribution of the number of Web domains presented to household members by the end of each day of the simulation. The box represents the IQR, the whiskers represent ±1.5× the IQR, the bold line represents the median of the distribution, and the dots represent outliers.

scheme, where only five Web pages are downloaded from each domain at a time before FiDO switches pushing content from the next ranked domain. We label this scheme as "User Pref. (RR)." The average daily domain coverage for the round robin user preference scheme is 0.20 ($\sigma = 0.23$). The reason the average daily domain coverage for the round robin prioritization scheme is less than what is provided by the push/pull scheme is because it provides more opportunities for content from domains prioritized by the surrounding community to be pushed to users whereas the push/pull approach is solely responsive to the specific Web browsing demands of a household. Thus, for the push/pull scheme, content is browsed from domains that households are interested in *on the day of the simulation*; the round robin user preference scheme browses content from a combination of domains that have historically been browsed by households and the domains most browsed by the community *on the day of the simulation.*

In addition to measuring daily domain coverage, we also measure the number of domains represented each day (see Figure 12). The round robin user preference scheme provides content from an average of 8.9 ($\sigma = 1.7$) different Web domains every day, which is 3.5 more domains than those provided by the push/pull scheme. In

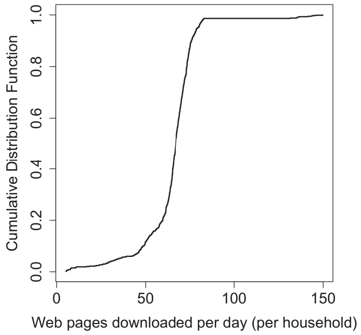


Fig. 13. Distribution of the number of Web pages provided to each household at the end of each day.

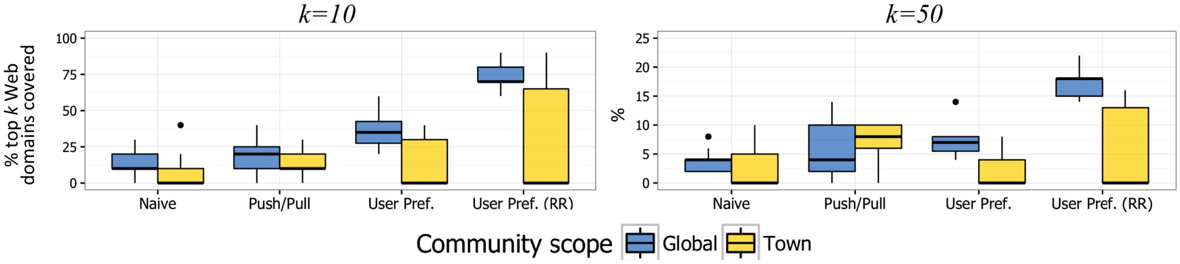


Fig. 14. Distribution of the percentage of the top *k* Web domains accessed by each household that are covered by FiDO using each prioritization scheme. The box represents the IQR, the whiskers represent ±1.5× the IQR, the bold line represents the median of the distribution, and the dots represent outliers.

Figure 13, we graph the distribution of the number of Web pages downloaded on behalf of each offline household during a single simulated day. On average, FiDO enables users to download 65 ($\sigma = 16$) Web pages on behalf of the members of their household each day.

In order to better understand how well FiDO is able to browse the Web on behalf of members of disconnected households, we measure the portion of the overall top k Web domains[4] that each prioritization scheme is able cover in each day of the simulation. We plot the percentage of the top 10 and top 50 Web domains that each prioritization scheme is able to cover in Figure 14. The round robin user preference approach covers the largest percentage of the top 10 ($\mu = 72.7\%$) and top 50 ($\mu = 17.2\%$) Web domains. We also examine the average rank of each of the Web domains crawled by the prioritization schemes in Figure 15. The rank corresponds inversely to the frequency with which the household accesses the domain during the overall observation period, so the ideal prioritization scheme would crawl domains with lower rankings. When examining the average rank of the Web domains crawled by the round robin user preference scheme we find the average rank is 10.4 ($\sigma = 6.2$), which is 9.6× smaller than the average rank of domains crawled by the push/pull scheme. We note that the user preference scheme is associated with the lowest average domain rank ($\mu = 2$; $\sigma = 2$), while it covers only a small percentage of the top 10 Web domains. This demonstrates how the addition of the round robin scheduling approach helps balance prioritization of the top ranked domains while also allocating resources across a broader range of domains.

[4] Based on the most accessed domains between January 17 and February 28, 2017.

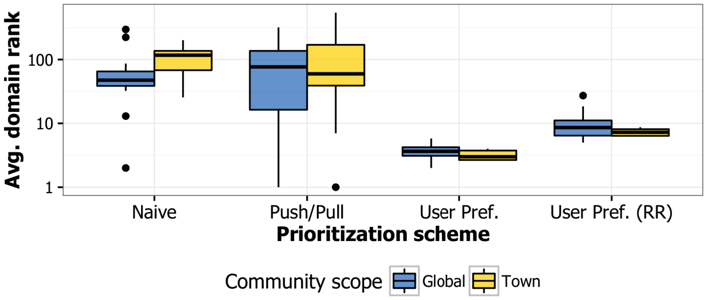


Fig. 15. Distribution of the average rank associated with Web domains that have pages pushed to users. Lower rank is better. The box represents the IQR, the whiskers represent ±1.5× the IQR, the bold line represents the median of the distribution, and the dots represent outliers.

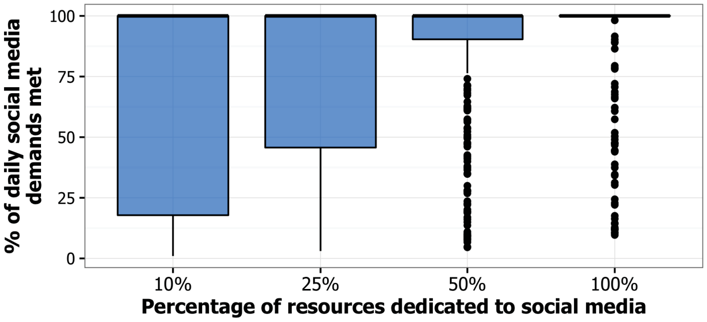


Fig. 16. Distribution of the percentage of households' daily social media demands that are met using the hybrid prioritization scheme. The box represents the IQR, the whiskers represent ±1.5× the IQR, the bold line represents the median of the distribution, and the dots represent outliers.

6.4.1 Hybridized prioritization. Related work [7, 20, 35, 58] as well as our own previous work [56, 57], demonstrate the importance of social media platforms for tribal communities. Social media platforms play a critical role in the tribal mediascape by empowering marginalized communities to take ownership of their representation in media, strengthen community bonds and notions of identity, and share cultural experiences and native language. In our analysis of Web traffic on the Red Spectrum network, we found that social media applications such as Facebook and YouTube were especially prevalent (see Section 4.2). Social media content poses a unique challenge to FiDO. Social media Web sites are extremely dynamic and highly dependent on the individual who is accessing. Social media is also prone to dynamic permissions policies, and as such, tokens or other authentication mechanisms are required to access social media content. These qualities make social media sites difficult to cache and browse with the community browsing and delivery paradigm with which FiDO operates. Nonetheless, we seek to alter FiDO operation to account for household social media usage. To do this, we introduce a hybrid approach, wherein some portion of a user's contact time with a base station is dedicated to downloading social media content on behalf of their household. We make two assumptions for this model: 1) there is a private and secure

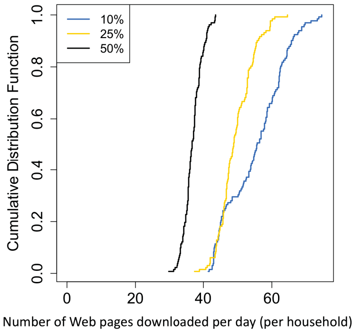


Fig. 17. Distribution of the number of web pages downloaded per day per household for the hybrid prioritization scheme operating with 10%, 25%, and 50% of resources dedicated to downloading social media content only.

way for individual social media users to share their authentication information with the household member who will be collecting content on their behalf and 2) a user's device can accurately predict its expected contact time with a base station.

We evaluate this hybrid approach to collecting social media content for disconnected households by combining the round robin user preference prioritization scheme with some percentage of contact time that is dedicated to downloading social media content on behalf of the household. It is imaginable that there are a multitude of configurations for this type of approach, for instance the percentage of the dedicated download resources that are allocated to each social media-consuming household member or the priority of certain social media sites over others. We simplify these different configuration to a model wherein a single block of an opportunity window is dedicated to all household members and all applications equally. Our evaluation of the hybrid approach involves measurement of the daily coverage and the portion of each household's aggregate daily social media needs (measured in bytes) covered by the hybrid approach. We evaluate FiDO operating with 10%, 25%, 50%, and 100% of its opportunity windows dedicated to downloading social media from Facebook, YouTube, Twitter, Instagram, and Snapchat.

In Figure 16, we plot the distribution of the percentage of each household's daily social media demands that are met with various configurations of the hybrid prioritization scheme. We find that even with only 10% of opportunistic resources allocated to downloading social media (an average of 46.9 MB per day), households are able to have an average of 64% of their daily social media download demands met, with 51% of households receiving all of their expected social media content. Additionally, when examining the number of Web pages downloaded on behalf of each household (see the distributions in Figure 17), we find with the 10% hybridization scheme, an average of 55.30 ($\sigma = 8.5$) Web pages are downloaded each day. This indicates the very real feasibility that the social networking needs of members of disconnected households can be adequately met opportunistically while also providing households with an ample volume of Web content for offline browsing.

7 DISCUSSION

While we focus our work on tribal reservations due to our current partnerships, our work is more broadly applicable to rural communities in general. Previous work by ourselves [29, 30, 55, 56] and others [21] demonstrates high locality of interest in multiple rural communities. This larger pattern of local interest and interest similarity suggests that the community-based predictive pre-fetching done by FiDO is applicable across a wide range of rural communities. Systems operation in rural, disconnected communities is non-trivial [5]. This makes simulation of operation in such an environment particularly challenging. There are two major challenges associated with simulating system usage in sparsely connected rural communities. First, there is a lack of data surrounding the mobility patterns associated with these users. We address this lack of mobility data by relying on census data about commuting habits (i.e., the amount of time spent commuting to work and the time of day when the commute is started) and employment status (i.e., the number of hours worked per week) of the community we study [52]. Another challenge with simulation of rural usage is the lack of high-fidelity coverage maps for wireless data rates in rural areas (particularly in areas with geographical features that interrupt line of sight connections). We address this in our simulation by relying on statistical models shaped around data rate information collected from Open Signal Map and statistics on mobile broadband connectivity [15, 37]. While our simulation simplifies some of the complexity of mobilization and connectivity through rugged and rural terrain, we believe that by evaluating FiDO in a trace-driven manner using conservative statistical models of connectivity, we are able to demonstrate that opportunistic content delivery coupled with community-driven browsing can be a successful way to bridge gaps in connectivity for areas that lack ubiquitous Internet access.

There are a number of concerns that arise when leveraging mobile users to collect their household content. One concern is the required storage capacity of the collection device. Based on our simulation environment, users collected an average of 55.3 MB per day. This means that users' devices (i.e., smartphones or tablets) must have allocated content storage prior to the start of their commute each day or have some way to offload content to a separate storage device. Furthermore, our simulation model assumes users can only connect opportunistically via cellular base stations that they encounter as part of their daily commute. However, in some scenarios users would have broadband access at their place of work or school (i.e., the final destination of their daily commute). FiDO could be extended to allow for users to take advantage of this broadband connectivity to fetch even more content on behalf of their household. This extension would require users to provision even more storage resources for fetched content or perform a second level of content prioritization as storage resources fill.

Similarly, once mobile users return to their households, they must share the content collected throughout the day with other members of the household. Future work would determine the proper user interface for sharing, likely either by uploading the day's content to a shared household content server, allowing individual devices to operate as local content browsers, or more simply, directly sharing the collection device with other household members [33]. In our hybrid model, we assume that members of disconnected households have a way to entrust access credentials and authentication tokens to commuting members of their community household. This model of entrusting people with information for delayed communication is common in delay tolerant networking [33, 51, 53]. Moreover, studies of mobile technology use in developing communities have revealed that actual usage (e.g., an entire family sharing a single smartphone) and information passing models required to support delay tolerant networking may not be compatible with current individual-oriented security and privacy paradigms used by most of the Web. There are two major challenges to the implementation of the hybrid prioritization scheme as presented in Section 6.4.1. First, by requesting content from providers (e.g., Facebook, Instagram, Netflix, YouTube) using a single IP address supporting multiple (on the order of tens or hundreds) authentication tokens, providers may interpret FiDO access patterns as an attack and deny service to the intermediary FiDO content delivery node. One way to address this issue in a deployment is to identify some of the top content platforms that would be accessed through the hybrid prioritization model and request a white-listed status for the IP address associated with intermediary FiDO content delivery nodes. A second challenge is that FiDO does not support secure content over TLS/SSL since it is unable to guarantee end-to-end connectivity in real-time. When we do implement FiDO, we can address this limitation by altering the hybrid prioritization model so that opportunistically connected devices collect secure content directly from the Internet without relying on FiDO content nodes as intermediaries. Ultimately, these limitations indicate that an important direction for future work is to design security and privacy mechanisms for communal content access models that depend on collaborative efforts between multiple individuals.

Future work seeks to deploy the FiDO system alongside one of the tribal-operated ISPs with which we partner (the Southern California Tribal Chairmen's Association Tribal Digital Village [48] and the Red Spectrum Communication Network). As with any deployment effort, there are unforeseeable challenges and complications that may arise that would effect the performance of FiDO [5]. Two particularly relevant factors that could effect general operation include changes in commute durations and routes and the achievable goodput data rate of Internet connectivity available along transportation corridors. A deployment effort would seek to address the issue of commute durations and routes by surveying potential participants and placing FiDO content delivery nodes alongside telecommunications infrastructure that coincides with the majority of users' typical commute routes, with some redundancy of content delivery node placement at infrastructure that occurs along the main alternative routes. It is important to note that since FiDO is designed for rural areas, we expect that there are only a limited number of routes that users could take to work, especially when traveling from clusters of homes that are located in remote areas. In addition to maximizing the value of node placement in a community, a deployment effort would involve signal measurement along some of the main routes taken by users. Obviously, as users drive closer to a cellular base station with line of sight connection, the goodput data rate is likely to increase; similarly, as routes meander through mountains and forested areas, the goodput data rate will change and likely decrease. FiDO addresses this intrinsically with the content prioritization schemes, which are inherently designed to select content for delivery over connections with low goodput. Additionally, FiDO content delivery nodes could monitor the data rate experienced by users' devices and alter the community scope to maximize the relevance of the content delivered over the connection. For instance, as we discuss in Section 4, usage patterns generated by a smaller geographic scope of community tends to more accurately predict content that would be relevant to a household if the connectivity opportunity window is small (either due to brief contact time or low data rates). While the work presented here is evaluated exclusively using trace-based simulations, we believe that this approach sufficiently evaluates the feasibility of the FiDO system under normal operation assumptions derived through observations made from our own usage data as well as commuter and connectivity data observed by transportation experts [14]. Indeed, rigorous, trace-driven simulations such as the one presented in this work are a critical part of ongoing collaborations between research institutions and tribal communities, wherein researchers can credibly demonstrate value, utility, and functionality of innovative systems to specific communities *prior* to requiring community partners to spend valuable time and resources deploying the system.

8 CONCLUSION

Web access is still far from ubiquitous and even in developed countries, pernicious digital divides persist [1, 16, 26, 53]. Our work seeks to ameliorate this divide by augmenting existing cellular infrastructure in a way that leverages community Web browsing similarities and opportunistic cellular connections. FiDO browses the Web on behalf of disconnected users by crawling the domains most accessed by the community and storing the crawled content at base stations located throughout the community. When users from disconnected homes mobilize through areas with mobile broadband availability, FiDO pushes the collected content to their device according to a prioritization scheme. In this paper, we seek to determine the feasibility of leveraging both community Web usage and opportunistic cellular connectivity in order to provide a Web browsing experience to users who live in areas where Internet access is not available.

Our analysis of Web traffic in a rural, Native American reservation demonstrates that the aggregate Web usage of a community can predict an average of 35% of any individual household's non-streaming, downloaded Web content and can predict 93% of the Web domains browsed by a household. Using trace-driven simulations and statistical models parameterized with data collected by the U.S. Census Bureau and Department of Transportation, we find that even with sparse connectivity available, an average of 80% of a household's cacheable Web files can be delivered opportunistically. Moreover, we find that when crawling the Web on behalf of disconnected households, FiDO is able to provide an average of 69.4 Web pages to each household (where 73% of a household's most browsed Web domains are represented by the content collected on their behalf). We further demonstrate how FiDO can accommodate both browsing and searching techniques using a hybrid prioritization scheme, wherein a certain percentage of download opportunities are dedicated to search tasks and the remainder are available to push browsed content. We evaluate this hybrid approach using requests for a user's social media feed as the search task; even with only 10% of opportunistic resources dedicated to downloading social media content, disconnected households receive an average of 64% of their daily social media content in addition to 55.3 Web pages that were fetched on their behalf. Critically, we demonstrate how FiDO can feasibly provide a Web browsing experience that navigates the online-offline transition characteristic of rural communities in a way that maximizes the value of existing information infrastructures.

9 ACKNOWLEDGEMENTS

This work was funded by the NSF Graduate Research Fellowship Program under Grant No. DGE-1144085 and grants NSF-1563436 and NSF-1637265. We would like to thank Valerie Fast Horse, Tom Jones, and Justin Hall at Red Spectrum Communications for their assistance in the collection of data used in this work.

REFERENCES

[1] V. Alia. *The New Media Nation: Indigenous Peoples and Global Communication*, volume 2. Berghahn Books, 2010.

[2] R. Armstrong, D. Freitag, T. Joachims, and T. Mitchell. WebWatcher: A Learning Apprentice for the World Wide Web. In *Proceedings of the AAAI Spring Symposium on Information Gathering from Heterogenous, Distributed Environments*, pages 59–66, Palo Alto, CA, USA, March 1995.

[3] R. Armstrong, D. Freitag, T. Joachims, and T. Mitchell. WebWatcher: A Learning Apprentice for the World Wide Web. In *AAAI Spring Symposium*, pages 6–12, Mar. 1995.

[4] A. Balasubramanian, Y. Zhou, W. Croft, B. Levine, and A. Venkataramani. Web Search from a Bus. In *Proceedings of the Second ACM Workshop on Challenged Networks*, pages 59–66, Montreal, Quebec, Canada, September 2007.

[5] E. Brewer, M. Demmer, M. Ho, R. Honicky, J. Pal, M. Plauche, and S. Surana. The Challenges of Technology Research for Developing Regions. *IEEE Pervasive Computing*, 5(2):15–23, 2006.

[6] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph Structure in the Web. *Computer Networks*, 33(1):309–320, 2000.

[7] P. Carpenter, K. Gibson, C. Kakekaspan, and S. OâĂŹDonnell. How Women in Remote and Rural First Nation Communities are Using Information and Communication Technologies (ICT). *Journal of Rural and Community Development*, 8(2), 2014.

[8] J. Carr. State Traffic and Speed Laws. http://www.mit.edu/~jfc/laws.html#types, April 2015.

[9] J. Chen, L. Subramanian, and J. Li. RuralCafe: Web Search in the Rural Developing World. In *ACM WWW 2009*, pages 411–420, Madrid, Spain, Apr. 2009.

[10] L. Chen and K. Sycara. WebMate: A Personal Agent for Browsing and Searching. In *Proceedings of the 18th International Conference on Autonomous Agents*, pages 132–139, Minneapolis, MN, USA, July 1998.

[11] L. Chen and K. Sycara. WebMate: A Personal Agent for Browsing and Searching. In *AAMAS 1998*, pages 132–139, Minneapolis, MN, USA, July 1998.

[12] M. D. Clark. *To Tweet Our Own Cause: A Mixed-methods Study of the Online Phenomenon "Black Twitter"*. The University of North Carolina at Chapel Hill, 2014.

[13] A. Clauset, C. R. Shalizi, and M. E. Newman. Power-law Distributions in Empirical Data. *SIAM Review*, 51(4):661–703, 2009.

[14] Department of Transportation. Bureau of Transportation Statistics. https://www.rita.dot.gov/bts/sites/rita.dot.gov.bts/files/publications/national_transportation_statistics/html/table_01_36.html, May 2016.

[15] H. S. Dhillon and J. G. Andrews. Downlink Rate Distribution in Heterogeneous Cellular Networks Under Generalized Cell Selection. *IEEE Wireless Communications Letters*, 3(1):42–45, 2014.

[16] M. E. Duarte. *Network Sovereignty: Understanding the Implications of Tribal Broadband Networks*. PhD thesis, University of Washington, 2013.

[17] R. Fagin, R. Kumar, and D. Sivakumar. Comparing Top k Lists. *Journal on Discrete Mathematics*, 17(1):134–160, 2003.

[18] K. Fall. A Delay-Tolerant Network Architecture for Challenged Internets. In *ACM SIGCOMM 2003*, pages 27–34, Berlin, Germany, Aug. 2003.

[19] Federal Communications Commission. Native Nations. https://www.fcc.gov/general/native-nations, Sept. 2016.

[20] K. Gibson, M. Kakekaspan, G. Kakekaspan, S. O'Donnell, B. Walmark, and B. Beaton. A History of Everyday Communication by Community Members of Fort Severn First Nation: From Hand Deliveries to Virtual Pokes. In *Proceedings of the 2012 iConference*, pages 105–111, 2012.

[21] E. Gilbert, K. Karahalios, and C. Sandvig. The Network in the Garden: An Empirical Analysis of Social Media in Rural Life. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, April, isbn = 978-1-60558-011-1, address = Florence, Italy, pages = 1603–1612 2008.

[22] S. Guo, M. H. Falaki, E. A. Oliver, S. Ur Rahman, A. Seth, M. A. Zaharia, and S. Keshav. Very Low-cost Internet Access Using KioskNet. *SIGCOMM Computer Communication Review*, 37(5):95–100, Oct. 2007.

[23] K. Heimerl and E. Brewer. The Village Base Station. In *ACM DEV 2010*, pages 131–140, San Francisco, CA, USA, June 2010.

[24] K. Heimerl, S. Hasan, K. Ali, E. Brewer, and T. Parikh. Local, Sustainable, Small-scale Cellular Networks. In *ICTD 2013*, pages 2–12, Cape Town, South Africa, Dec. 2013.

[25] HTTP Archive. Average Bytes per Page by Content Type. http://httparchive.org/interesting.php?a=All&l=Apr%201%202017, April 2017.

[26] International Telecommunications Union. Facts and Figures. http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2016.pdf, June 2016.

[27] S. Isaacman, R. Becker, R. Cáceres, M. Martonosi, J. Rowland, A. Varshavsky, and W. Willinger. Human Mobility Modeling at Metropolitan Scales. In *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*, MobiSys '12, pages 239–252, Low Wood Bay, Lake District, UK, June 2012.

[28] S. Jain, K. Fall, and R. Patra. *Routing in a Delay Tolerant Network*, volume 34. 2004.

[29] D. Johnson, E. Belding, and G. Van Stam. Network Traffic Locality in a Rural African Village. In *ICTD 2012*, pages 268–277, Atlanta, GA, USA, Mar. 2012.

[30] D. Johnson, V. Pejovic, E. Belding, and G. van Stam. VillageShare: Facilitating Content Generation and Sharing in Rural Networks. In *ACM DEV 2012*, pages 61–70, Atlanta, GA, USA, Mar. 2012.

[31] T. T. Keegan, P. Mato, S. Ruru, et al. Using Twitter in an Indigenous Language: An Analysis of Te Reo Maori Tweets. *AlterNative: An International Journal of Indigenous Peoples*, 11(1):59, 2015.

[32] H. Lieberman. Letizia: An Agent that Assists Web Browsing. In *Proceedings of the 1995 International Joint Conference on Artificial Intelligence*, pages 924–929, Montreal, Quebec, Canada, August 1995.

[33] T. Matthews, K. Liao, A. Turner, M. Berkovich, R. Reeder, and S. Consolvo. She'll just grab any device that's closer: A study of everyday device & account sharing in households. In *CHI 2016*, pages 5921–5932, San Jose, CA, USA, September 2016.

[34] R. Meusel, S. Vigna, O. Lehmberg, and C. Bizer. Graph Structure in the Web—Revisited: A Trick of the Heavy Tail. In *Proceedings of the 23rd international conference on World Wide Web*, pages 427–432, Seoul, South Korea, May 2014.

[35] H. Molyneaux, S. O'Donnell, C. Kakekaspan, B. Walmark, P. Budka, and K. Gibson. Social Media in Remote First Nation Communities. *Canadian Journal of Communication*, 39(2), 2014.

[36] National Telecommunications and Information Administration. Exploring the Digital Nation: Embracing the Mobile Internet, October 2014.

[37] Open Signal Map. Compare Mobile Networks Near You. https://opensignal.com/, March 2017.

[38] J. Ott and D. Kutscher. A Disconnection-tolerant Transport for Drive-thru Internet Environments. In *Proceedings of the 24th Annual Joint Conference of the IEEE Computer and Communications Societies*, volume 3, pages 1849–1862, Miami, FL, USA, March 2005.

[39] M. Pazzani. A Framework for Collaborative, Content-based and Demographic Filtering. *Artificial Intelligence Review*, 13(5-6):393–408, 1999.

[40] M. Pazzani, J. Muramatsu, and D. Billsus. Syskill & Webert: Identifying Interesting Web Sites. In *Proceedings of the 13th National Conference on Artificial Intelligence*, pages 54–61, Portland, OR, USA, August 1996.

[41] A. Pentland, R. Fletcher, and A. Hasson. DakNet: Rethinking Connectivity in Developing Nations. *Computer*, 37(1):78–83, 2004.

[42] M. Pitkänen and J. Ott. Enabling Opportunistic Storage for Mobile DTNs. *Pervasive and Mobile Computing*, 4(5):579–594, 2004.

[43] S. C. Rushing and D. Stephens. Use of Media Technologies by Native American Teens and Young Adults in the Pacific Northwest: Exploring Their Utility for Designing Culturally Appropriate Technology-based Health Interventions. *The Journal of Primary Prevention*, 32(3):135, 2011.

[44] J. Schafer, D. Frankowski, J. Herlocker, and S. Sen. Collaborative Filtering Recommender Systems. In *The Adaptive Web*, pages 291–324. Springer, 2007.

[45] P. Schmitt, R. Raghavendra, and E. Belding. Internet Media Upload Caching for Poorly-connected Regions. In *ACM DEV 2015*, pages 41–49, London, UK, Dec. 2015.

[46] A. Seth, D. Kroeker, M. Zaharia, S. Guo, and S. Keshav. Low-cost Communication for Rural Internet Kiosks Using Mechanical Backhaul. In *ACM MobiCom 2006*, pages 334–345, Los Angeles, CA, USA, Sept. 2006.

[47] A. Smith. U.S. Smartphone Use in 2015. http://www.pewinternet.org/2015/04/01/us-smartphone-use-in-2015/, April 2015.

[48] Southern California Tribal Chairmen's Association. Tribal Digital Village. https://sctdv.net/, January 2017.

[49] The Bro Project. Bro Network Security Monitor. https://www.bro.org/, Sept. 2016.

[50] TraceAnon. http://www.wand.net.nz/trac/libtrace/wiki/TraceAnon, July 2010.

[51] C. A. Trujillo, A. Barrios, S. M. Camacho, and J. A. Rosa. Low Socioeconomic Class and Consumer Complexity Expectations for New Product Technology. *Journal of Business Research*, 63(6):538 –547, 2010.

[52] United States Census Bureau. American Fact Finder. http://factfinder.census.gov, 2014.

[53] T. Unwin. *ICT4D: Information and Communication Technology for Development.* Cambridge University Press, 2009.

[54] M. Vigil, E. Belding, and R. M. Repurposing FM: Radio Nowhere to OSNs Everywhere. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 1260–1272, San Francisco, CA, USA, February 2016.

[55] M. Vigil, E. Belding, and M. Rantanen. Repurposing FM: Radio Nowhere to OSNs Everywhere. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 1260–1272, San Francisco, CA, USA, 2016.

[56] M. Vigil, M. Rantanen, and E. Belding. A Fist Look at Tribal Web Traffic. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1155–1165, Florence, Italy, May 2015.

[57] M. Vigil-Hayes, M. Duarte, N. D. Parkhurst, and E. Belding. *#indigenous*: Tracking the Connective Actions of Native American Advocates on Twitter. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1387–1399, Portland, OR, USA, 2017.

[58] J. Waitoa, R. Scheyvens, T. R. Warren, et al. E-whanaungatanga: The Role of Social Media in Maori Political Empowerment. *AlterNative: An International Journal of Indigenous Peoples*, 11(1):45, 2015.

[59] W. Zhao, M. Ammar, and E. Zegura. A Message Ferrying Approach for Data Delivery in Sparse Mobile Ad Hoc Networks. In *ACM MobiHoc 2004*, pages 187–198, Roppongi, Japan, May 2004.

[60] M. Zheleva, A. Paul, D. Johnson, and E. Belding. Kwiizya: Local Cellular Network Services in Remote Areas. In *ACM MobiSys 2013*, pages 417–430, Taipei, Taiwan, June 2013.