University of California
Santa Barbara

Community-based Networks for Challenged Environments

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Computer Science

by

Morgan Vigil-Hayes

Committee in charge:

Professor Elizabeth Belding (Advisor), Chair
Professor Amr El Abbadi(Committee Member)
Professor Ben Zhao (Committee Member)
Professor Ellen Zegura (Committee Member)

June 2017

The Dissertation of Morgan Vigil-Hayes is approved.

Professor Amr El Abbadi(Committee Member)

Professor Ben Zhao (Committee Member)

Professor Ellen Zegura (Committee Member)

Professor Elizabeth Belding (Advisor), Committee Chair

May 2017

Community-based Networks for Challenged Environments

Morgan Vigil-Hayes

For all my relations.

Acknowledgements

This work is the culmination of years of blood, sweat, and tears shed amongst loved-ones and mentors who contributed to the smiles and laughter that made the completion of this dissertation possible.

First and foremost, I would like to thank my husband, Isaac, for joining me in this journey and walking patiently alongside me. You saw me at my best and worst during this process of learning, and you always believed in me and supported my vision.

Significant thanks go to my research advisor, Elizabeth Belding. I cannot begin to list all the ways in which you helped me grow as an explorer, creator, and sharer of knowledge. I am grateful for your patient mentoring and constant encouragement. I am thankful for the researcher and person I have become after spending five years under your guidance.

Thanks are also due to my committee members, Ben Zhao, Amr El Abbadi, and Ellen Zegura. Your feedback and encouragement has been invaluable to this process. You have been excellent teachers in evaluating research and identifying new possibilities and opportunities. I also had the great privilege to collaborate with Ellen and it was a joy to work with her to create new ideas.

To my labmates and collaborators: Paul Schmitt, Waylon Brunette, Mai El Sherieff, Danny Iland, Mariya Zheleva, Michael Nekrasov, and David Johnson; you understood the process like no one else, provided the earliest feedback, and shared many moments of laughter. Your diligence, thoughtfulness, and helpfulness have inspired and motivated me over the course of this dissertation. I especially want to thank Paul, who was my perpetual project partner for the first two years of graduate school and put up with my antics and propensity to lose keys with patience and good humor.

To the American Indian academic community: Marisa Duarte, Margaret McMurtrey,

Keri Bradford, and Linda Murray; your work in academia, minds for problem solving, and hearts for the people around you have served as a model for me. In particular, I would like to thank Marisa, who encouraged and mentored me when I took a risk and went down a path of study that was nontraditional. Your dissertation was the first I read and it is an honor that I can acknowledge you in mine.

This work would not have been possible without the love and support of friends and family. Wayne counseled me through the rollercoaster of graduate school and constantly affirmed my place in the academic community. Kristen always asked about my work and understood how much it was a part of me. Stacey and T.J. made sure I took breaks for good food, hot tea, and fun games. Danielle distracted my mind with her many antics ("and how!"). Josette was my wise house elf. Jon took me to Spain and inspired me to get outside and explore. Most importantly, my parents, Ralph and Stephanie, raised me to be curious, caring, and resilient. They taught me to love deeply and to pursue my vision, even when failure and doubts made it challenging. Mom and Dad, I cannot begin to express just how much you are my heroes. Thank you for teaching me how to ask many questions and pray thoughtfully.

Finally, I would like to acknowledge significant mentors who passed away during the completion of this work. Kim Kihlstrom's early words of encouragement helped keep me brave when facing the new and unknown. Her hospitality to students will always be an inspiration to me. Gaetano Borriello's caring spirit and sharp mind inspired me to think critically about what the world actually needs.

Curriculum Vitæ
Morgan Vigil-Hayes

Education

2017	Ph.D. in Computer Science (Expected), University of California, Santa Barbara.
2017	M.S. in Computer Science, University of California, Santa Barbara.
2011	B.S. in Computer Science, Westmont College. *summa cum laud*

Awards

University of California, Santa Barbara

Doctoral Scholars Fellowship, 2012–2017

NSF Graduate Research Fellowship Program Fellowship, 2013–2016

Microsoft Graduate Women's Scholarship, 2013

N2Women Fellowship for HotMobile, 2017

Honorable Mention for Best Paper Award at CSCW 2017

Westmont College

Grace Hopper Award for Top Computer Science Graduate, 2012

David K. Winter Servant Leader Award, 2011

Presidential Scholar, 2008–2011

NSF Computer Science Scholarship, 2008–2011

Publications

1. **Vigil-Hayes, M.**, Belding, E., Zegura, E. "FiDO: A Community-based CDN for Challenged Network Environments," *in submission*, 2017.
2. **Vigil-Hayes, M.**, Duarte, M., Parkhurst, N.D., Belding, E., "*#Indigenous*: Tracking the Connective Actions of Native American Advocates on Twitter," *CSCW 2017*, Portland, OR, USA. March 2017.
3. Schmitt, P., **Vigil, M.**, Belding, E., "A Study of MVNO Data Paths and Performance," *PAM 2016*, Heraklion, Crete, Greece. March 2016.
4. **Vigil, M.**, Belding, E., Rantanen, M., "Repurposing FM: Radio Nowhere to OSNs Everywhere," *CSCW 2016*, San Francisco, CA, USA. March 2016.

5. Brunette, W., **Vigil, M.**, Pervaiz, F., Levari, S., Borriello, G., Anderson, R. "Optimizing Mobile Application Communication for Challenged Network Environments," *ACM DEV 2015*, London, England, UK. December 2015.
6. Zheleva, M., Schmitt, P., **Vigil, M.**, Belding, E. "Internet Bandwidth Upgrade: Implications on Performance and Usage in Rural Zambia." *Information Technologies & International Development (ITID)*. 11:2, 1–18. Spring 2015.
7. **Vigil, M.**, Rantanen, M., Belding, E., "A First Look At Tribal Web Traffic." *WWW '15*, Florence, Italy. May 2015.
8. Zheleva, M., Schmitt, P., **Vigil, M.**, Belding, E., "The Increased Bandwidth Fallacy." *ACM DEV '13*, Cape Town, South Africa. December 2013.
9. Zheleva, M., Schmitt, P., **Vigil, M.**, Belding, E., "Bringing Visibility to Rural Users in Ivory Coast." *ICTD '13*, Cape Town, South Africa. December 2013.
10. Zheleva, M., Schmitt, P., **Vigil, M.**, Belding, E., "Community Detection in Cellular Network Traces." *ICTD '13*, Cape Town, South Africa. December 2013.
11. Schmitt, P., Zheleva, M., **Vigil, M.**, Belding, E., "Communication Flow Patterns in the Orange Telecom D4D Datase." *NetMob '13*, Boston, MA. January 2013.

Experience

Research Assistant — September 2012 to present

Department of Computer Science,
University of California
Supervisor: Elizabeth Belding, Ph.D.

Research Assistant — June 2013 to August 2013

Department of Computer Science,
University of Washington
Supervisor: Gaetano Borriello, Ph.D.

Instructor — Summer 2015

CMPSC 16 - Problem Solving with Computers (C++)
Department of Computer Science,
University of California, Santa Barbara

Teaching Assistant — 2009–2010

CS 10 - Introduction to Programming II (Scheme)
Instructor: Wayne Iba, Ph.D.
Department of Computer Science,
Westmont College

CS 30 - Introduction to Programming III (Ruby)
Instructor: Kim Kihlstrom, Ph.D.
Department of Computer Science,
Westmont College

CS 45 - Computer Architecture and Organization
Instructor: Kim Kihlstrom, Ph.D.
Department of Computer Science,
Westmont College

Abstract

Community-based Networks for Challenged Environments

by

Morgan Vigil-Hayes

The Internet as a networked system has been rendered more complex than ever before as human endpoints are grafted into the system via increasingly pervasive and personalized networked devices. According to the United Nations, the Internet is a transnational enabler of a number of human rights, and as such, access to the Internet has been proclaimed to be a basic right unto itself. Unfortunately, even as networked devices have become ubiquitous, access to the Internet has not. In many cases, the reasons behind this digital divide involve contextual challenges such as limited infrastructure, limited economic viability, and rugged terrain. In this dissertation, we seek to ameliorate these challenges by designing data-driven, community-based network infrastructure.

In order to extend Internet connectivity to communities located in some of the most challenging contexts, we start by understanding how Internet connectivity is used when communities receive initial Internet access. We do this by partnering with two ISPs (Internet service providers) that brought initial Internet connectivity to two geographic regions in Indian Country. The data we have collected from these two ISPs totals to 115 TB generated over a combined three years of partnerships. Our ISP collaborators serve a total of 1,300 subscribers who represent residents of 14 different Native American reservations representing 18 different tribes. The service areas of these ISPs include predominantly rural communities located on mountainous and forested terrain. Key findings from our analysis of data generated by these ISPs include: the prevalence of social media and streaming content, the locality of interest with respect to social media

content, and the similarity of Web browsing preferences between households and the aggregate communities to which they belong. We augment our analysis of network traces collected from ISPs with analysis of data collected from some of the most prevalent social media platforms. One of our studies mines Instagram trace data collected from Instagram servers to better understand the relationship between network infrastructure capacity and social media usage patterns. We found that only a small percentage of content available to users over social media platforms is actually interacted with by users and that only a small portion of available bandwidth is needed to support interaction with this content. Moreover, in our analysis of the diffusion of content disseminated by Native American advocates on Twitter, we found that the rate of diffusion and the prevalence of content is tied to its media richness, and that richer content does not guarantee rapid diffusion or longevity in the network. Based on the results our analyses as well as findings in related work, we design four community-based network technologies that address the network challenges associated with rural and developing contexts.

First, we introduce a social media content distribution system that operates over FM radio [200]. In order to provide content over a 1.2 Kbps technology (the Radio Broadcast Data System), we create a graph-based metric, the cumulative clustering coefficient, to filter content based on its total audience size and the diversity of its audience scope. We evaluate this delivery system used a trace-based simulation and we find that 81% of users received at least half of their content requests and 35.5% of the 1.1 million requested Instagram photos were transmitted to users. Next, we introduce FiDO [203], a community-based Web browsing agent and content delivery system that enables users from disconnected households to collect relevant content for themselves and members of their households opportunistically from content caches co-located with cellular base stations. We evaluate FiDO using a trace-driven simulation that combines Web traces collected from one of our partner ISPs in addition to statistical models parameterized

with census and transportation data. We find that an average of 80% of a household's cacheable Web files can be delivered opportunistically and when crawling the Web on behalf of disconnected households, FiDO is able to provide an average of 69 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 then describe some of the challenges associated with content creation and data collection in challenging contexts and introduce Open Development Kit (ODK) Submit and VillageShare for rural schools. ODK Submit is a smartphone-based platform that sits between data collection applications and the network interfaces of a devices [26]. It seeks to ease the burden of navigating heterogeneous network conditions for application developers, data collectors, and data processors. Principles from ODK Submit were incorporated into the publicly available ODK v. 2.0 tool suite as part of the Aggregate Tables Extensions suite [143]. In addition, we introduce VillageShare for rural schools, which enables schools in poorly-connected, rural areas to create and share culturally relevant curricula and empowers students to work collaboratively on "local cloud-based" projects despite their lack of access to network connectivity at home. We provide an evaluation of VillageShare that has been informed and parameterized by the deployment of Internet connectivity to rural schools over high-latency, low-bandwdith technology in South Africa.

We conclude with an overview of our key findings as well as a discussion of future research directions inspired by the work in this dissertation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The Internet as a networked system has been rendered more complex than ever before as human endpoints are grafted into the system via increasingly pervasive and personalized networked devices. According to the United Nations, the Internet is a transnational enabler of a number of human rights, and as such, access to the Internet has been proclaimed to be a basic right unto itself [113]. Unfortunately, even as networked devices have become ubiquitous, access to the Internet has not.

Investigations into the global penetration of Internet access have revealed access divides that are stark as well as subtle. By the beginning of 2017, 53% of the world's population continued to lack access to the Internet[1]; most affected were the least developed countries, where 85% of people did not have access to the Internet [99]. When examining the quality of Internet access, more discrepancies become apparent. When measuring the average data rate available to Internet users in different countries, the International Telecommunications Union revealed that in developing countries, the majority of fixed broadband speeds were between 256 Kbps and 2 Mbps [98].

In the past decade, the most dynamic broadband technology with respect to global

[1]This includes access via all technologies including fixed-broadband and mobile-broadband.

subscription growth rate has been mobile broadband which reached a penetration rate of 47% in 2015 from 4% in 2007 [98]. However, even mobile broadband deployment has not bridged access gaps equally in all contexts and pernicious digital divides persist. In rural areas, only 29% of people have access to 3G mobile-broadband, whereas in urban areas, 89% of people have access to the same technology. Typically, the reasons behind rural-urban digital divides include: sparse population density (which equates to limited market size) and distance from existing core infrastructures. This disparity is extremely problematic in developed countries where societal expectations of broadband accessibility clash with the reality of rural-urban digital divides. When "plug-and-play" wireless broadband infrastructure is touted as a panacea to rural connectivity [98, 66], it is crucial to critically evaluate these claims in the face of limited purchasing power in rural areas [10], the rapid growth in Web content sizes and the increasingly interactive nature of content delivered over the Internet [91, 132]. Thus, while the cost of a pre-paid 1 GB mobile-broadband subscription is at an all-time low [98], the "content purchasing power" of that bandwidth capacity has decreased; moreover, the attainable data rates of cellular networks leave much to be desired by way of end-user experience for the demands of increasingly interactive applications [148, 169, 33].

Unfortunately, this pattern of Internet divide promises to persist [80]. As new ICTs are developed and as new applications push the cutting edge in responsiveness and real-time interaction, broadband infrastructures that once functioned as an initial step across the digital divide are positioned further away from edge of the gap. However, this disparity might be moderated by expanding our notion of technologies that bridge digital divides. To date, the majority of efforts to bridge digital divides involve building out new infrastructure that provides new and better broadband access to poorly connected communities. In this dissertation, we argue that this existing concept of bridge technologies should be expanded to include technologies that augment existing broadband

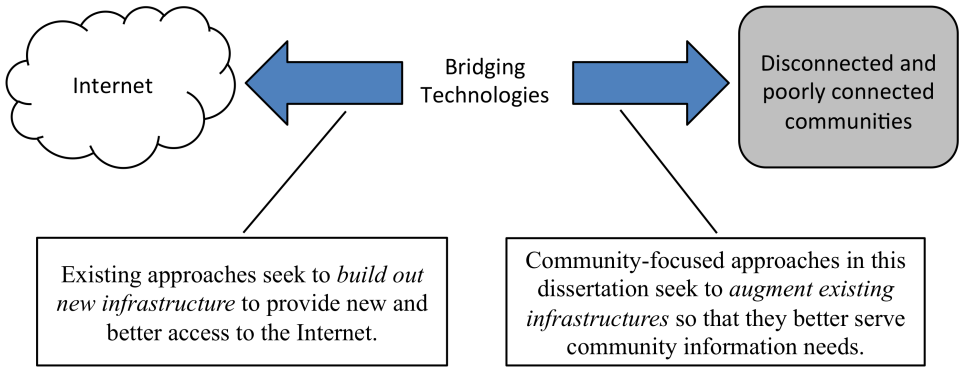


Figure 1.1: Existing approaches to bridging the digital divide focus on building out new infrastructures whereas the approaches discussed in this dissertation seek to expand the notion of bridging technologies to include systems that augment and add value to existing infrastructures.

connectivity in such a way that its value with respect to information delivery capacity is increased. We illustrate how these two different approaches compliment each other in efforts to minimize digital divide effects in Figure 1.1

We refer to these community-centered augmentations as "community-based networks." In order to understand and evaluate community-based networks, it is necessary to partner with communities. We focus on partnering with two communities in this work: the Southern California Tribal Chairmen's Association Tribal Digital Village (TDV) network and the Coeur d'Alene Indian Tribe's Red Spectrum Communications network. As networks operating in communities in Indian Country, the networks we partner with provide unique insight into the usage behaviors of communities where access to Internet connectivity is not universal (i.e., some areas have access and some do not; some areas with access have higher capacity access than others). In order to provide readers with a full appreciation of the implications of designing ICTs with partners from Indian Country, we provide an extensive background on the context of our collaborators in Chapter 2.

1.1 Thesis and Contributions

This dissertation demonstrates that:

> *Community-based networks deployed in challenged environments can be more effectively designed and utilized if they take advantage of the social, geographical, and topical communities that users form independently of the network infrastructure.*

The overview of this dissertation is depicted in Figure 1.2. We utilize a two-step approach to designing community-based network technologies. First, we collect data about community network usage patterns and analyze that data to identify opportunities for optimizing content delivery over space and time. Second, we design content delivery systems based on our observations of community usage and community structures, which we evaluate using trace-driven simulations. We provide a short summary of the work associated with each category outlined in Figure 1.2 as well as how the work contributes to the state of the art.

Network Characterization. We characterize networks and the communities that utilize them with respect to social connectivity, geographical relationships, and content interests.

- ***Characterizing Web traffic on reservations.*** Our work is the first to take a network analytic approach to the characterization of Web usage in Indian Country in the U.S. [202]. From our 115 TB of collected traffic, we mined trace data generated by 29.52 TB of traffic collected from the TDV network between June 2014 and August 2014. We identify which applications are most pervasive and we connect application usage with network performance. We are able to mine a social network from the URL fields of packet headers we collect from the TDV network

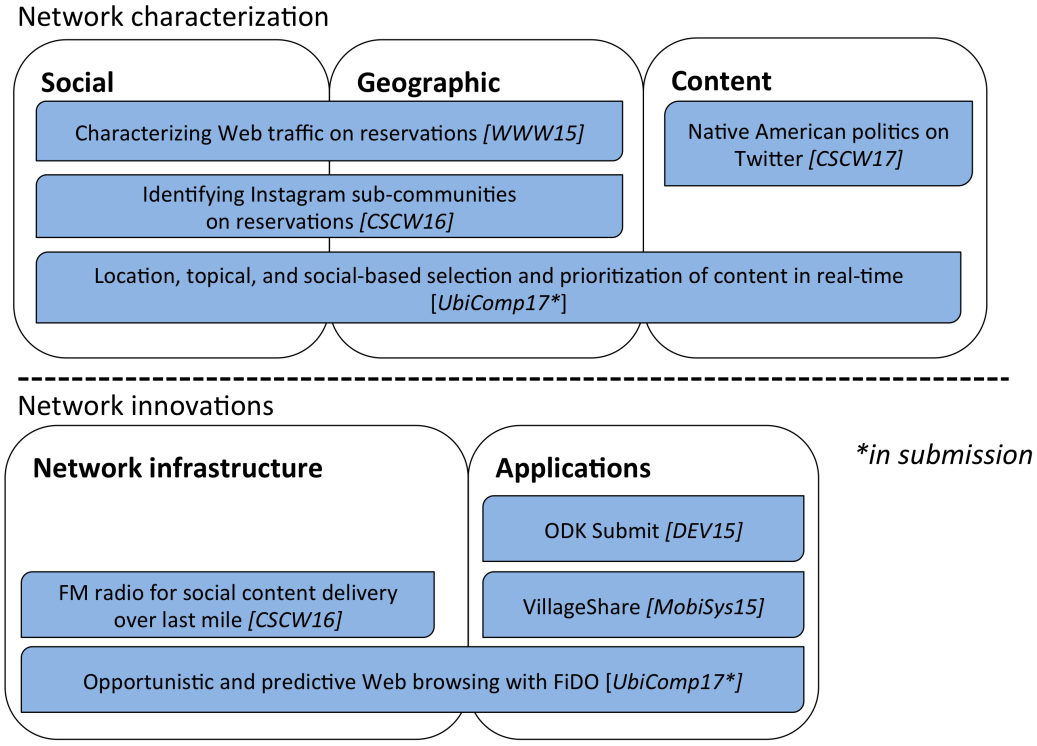


Figure 1.2: Dissertation overview.

and through analysis of these usage patterns, we demonstrate a high locality of interest wherein users in the TDV network are far more likely to interact with content generated by users from the same reservation as opposed to users from outside the same reservation. We also demonstrate how locally created social media content circulates in the network for a longer period of time and receives more interactions per media object. This work was central to characterizing how Internet connectivity aids in the strengthening of community bonds, which is critical for underrepresented cultures and marginalized communities.

- ***Identifying Instagram sub-communities on reservations.*** We extend our work characterizing Web traffic on reservations to incorporate usage trace data collected from Instagram servers using the Instagram API seeded with meta-data mined from the TDV Web traces [200]. We combine the Web traces with Instagram

traces collected over a six month period and provide insight into the similarity between the social networks of users in the TDV network, the rates at which TDV users interact with Instagram content, and the distinction between content that receives a high volume of interaction from TDV Instagram users and content that receives no interactions (despite its availability and the automatic download of that content). Critically, we find that only a small portion (0.55%) of available social media content is actually interacted with, and if we were to filter only content user's interact with, the necessary data rate to support the delivery of that relevant content would be 0.195 Kbps.

- ***Native American politics on Twitter.*** While our work largely focuses on the social media interactions between users in specific, geographically defined reservation communities, we wanted to understand community dynamics on a larger scale. Specifically, we are interested in understanding how information diffuses through the complex, interconnected social networks of a marginalized, minority community. To approach these questions, we seek to understand how Native Americans distribute and engage with political information on Twitter. Given the heterogeneity of Native American peoples in the U.S., we attempt a holistic understanding of Native American political discourse on social media by characterizing how Native American advocates utilize social media platforms for connective action. Using a post-structural, interdisciplinary, mixed methods approach, we use theories of connective action [15] and media richness [116] to analyze a Twitter data set culled from influential Native American advocates and their followers during the 2016 primary presidential election season. Our study sheds light on how Native American advocates use social media to propagate political information and identifies issues that are central to the political discourse of Native American advocates.

Furthermore, we demonstrate how the bandwidth characteristics of content impact its propagation and we discuss this in the context of pernicious digital divide effects present in Indian Country.

- ***Location, topical, and social-based selection and prioritization of content in real-time.*** In addition to examining usage similarity patterns of social media in a tribal-operated network, we examine the extent to which Web preferences reflected highly localized patterns of content interest. Our analysis of Web traffic in the Red Spectrum Communications network demonstrated 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. This finding was critical to establishing the feasibility of creating infrastructure that leverages community usage as a means to apply collaborative filtering mechanisms to prioritize bandwidth resources [203].

Network Innovation. We design innovative network infrastructures and applications that augment existing network infrastructure and ameliorate the movement between areas with high bandwidth Internet connectivity and areas with poor or no Internet connectivity.

- ***Repurposing FM radio.*** Inspired by our findings of the importance of social media sharing, overlapping social media content interests, and the lack of ubiquitous Internet connectivity on tribal lands, we propose a social media content delivery system that prioritizes and broadcasts the most relevant content via the radio broadcast data system (RBDS), which is the data subcarrier of FM radio. The proposed system leverages the advantages of RBDS technology: it propagates data over long distances, it is robust to transmission errors, and FM infrastructure is already ubiquitous on tribal lands. Since RBDS is an extremely low-bandwidth

broadcast technology, we proposed a content scheduling algorithm which leverages users social connectivity and round robin scheduling in order to allow for fair sharing of the broadcast medium. We create a metric to identify nodes with the widest and most embedded audience and we introduced an algorithm that prioritizes content based on the cumulative clustering coefficient associated with the creators of content. Using six months of Instagram traces generated by TDV users, we evaluate the fairness and coverage provided by the system. Our scheduling algorithm was able to provide half of the users in the community with 81% of their Instagram content requests and 35.5% of the 1.1 million requested Instagram photos were delivered to users over the six month simulation period. The key finding of this work is that community social networks can be leveraged to select relevant content for distribution over extremely low bandwidth [200].

- ***FiDO.*** Based on our findings of similarity in the Web browsing preferences within reservation communities, we propose FiDO, a community-based CDN that delivers content opportunistically [203]. FiDO is designed specifically to serve the households lacking access to the Internet by delivering content to mobile users on behalf of their entire household. FiDO adds value to existing cellular infrastructure located along traffic corridors by fetching content that members of disconnected household are likely to be interested in viewing (based on community usage patterns) and storing that content on servers co-located with cellular base stations. As users drive to work or school as part of their daily commute, their devices establish opportunistic connections to the base stations and download the curated content on behalf of the users and the members of their households. We evaluate the feasibility that FiDO has to deliver all of a household's daily Web content opportunistically as well as the extent to which community usage was predictive of household content needs.

Our trace-driven evaluations reveal 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. We further demonstrate that FiDO can accommodate both browsing and searching techniques using a hybrid prioritization scheme, and that 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.

- ***ODK Submit.*** Mechanisms that abstract the heterogeneity of Internet connectivity (i.e., connected vs. disconnected, well-connected vs. poorly connected) are critical to the operation of mobile data collection applications. Specifically, the design of mobile applications for data collection in challenged network environments necessitates new abstractions that target deployment architects, and non-developers who are charged with adapting an ensemble of off-the-shelf software to a deployment context. Data transfer is integral to mobile application design and deployments have inherent and contextual requirements that determine what data should be transferred and when. To this end, we introduce a new software tool called ODK Submit to help streamline application customization to challenged network environments. Submit is an Android service that coordinates data communication by providing channel monitoring and transmission scheduling mechanisms to Android apps. Submit provides software developers with an interface that abstracts communication channels and flexibly handles data ownership and application-specific synchronization issues. Critically, Submit introduces the concept of separating data semantics from physical data characteristics so that data transmission resources can

be utilized in a way that aligns with an organization's data collection values and objectives [26].

- ***VillageShare.*** Designed in response to observations of high rates of locality of interest in a rural network in Zambia [106, 107, 108], VillageShare functions as a community content cache that enables users in a community to share media files with each other via a local cloud server. We extend the single-server design of VillageShare to enable synchronization between multiple VillageShare servers [201]. Our multi-instance design of VillageShare was designed to facilitate collaboration between schools in rural and developing contexts where culturally relevant curricula can be scarce and collaborative work is hindered by the fact that Internet connectivity may be poor at a school (for instance, the school is the gateway access point for the entire community and connects to the Internet over a lossy, 1 Mbps satellite connection) and the fact that many students in these contexts do not have access to the Internet at home. Our design distinguishes between local content collaborations and non-local collaborations, which helps maximize bandwidth use. The VillageShare application allows users to work on collaborations offline before synchronizing with VillageShare servers when connectivity can be established.

1.2 Intellectual Impact

Here, we present the major intellectual impact of our work divided into two areas: network analysis and network system design.

1.2.1 Network Analysis and Characterization

Our network characterization studies make several important contributions to the areas of Computer Supported Cooperative Work (CSCW), Information and Communication Technologies for Development (ICT4D), and network analytics, both with respect to our findings as well as to our methodologies. First, as part of our work, we collected a total of 115 TB of network traffic traces from two rural, tribal-operated networks. Our analyses of usage behaviors in these networks were the first characterizations of network traffic in the context of Indian Country and they were the first studies to highlight distinctive usage behaviors in this context. We demonstrate a high locality of interest, wherein users are much more likely to interact with content generated by users from their own local community (i.e., reservation) than content generated in other reservations or content generated outside their network [202, 200]. We also demonstrate how socioeconomic realities impact network usage in Indian Country; e-commerce traffic that is highly prevalent in networks used by the general U.S. population is largely absent from the TDV and Red Spectrum networks [202]. This highlights how general lack of infrastructure as well as institutional barriers can dampen the effect Internet connectivity has on economic opportunity. A second contribution is our characterization of locality of interest amongst geographically-near Internet users. We demonstrate that users are more likely to interact socially with other users from the same community (as opposed to users in communities nearby) [202, 200] and user Web browsing preferences are more closely related to those of users from the same community [203]. Third, our work is the first to quantitatively confirm the prevalence of social media traffic in Indian Country, and specifically, we are the first to quantitatively confirm the role of these applications in strengthening community bonds as evidenced by the strong locality of interest that we observe. These observational findings play an important role in understanding how

network technologies can be re-designed in poorly-connected communities in order to increase the value of existing ICT infrastructure.

Our second set of contributions comes from our innovative methodological approaches to synthesizing disparate data sets to provide a more comprehensive perspective on the relationship between computer network systems and social and content networks that users form. In our work analyzing Instagram usage in the TDV network, we use data mined from packet headers to collect usage data from Instagram servers using the Instagram API. By matching information that is only accessible via the Instagram trace data to information that is only accessible via the TDV network traces, we generate a compound data set that provides a holistic perspective on the usage of a social network platform in a tribal community. This compound data set provided insight into information pertinent to computer networks such as the sizes of social media posts generated by a composition of individual objects, media meta-data that could be cross-referenced to media downloads in the traces and interactions with certain media objects. The compound data set also provided insight into some of the geographic characteristics of social networks by mapping user and media identifiers to specific geographic communities. Ultimately, the unique insights from this compound data set enabled us create algorithms that took advantage of network characteristics in such a way that infrastructure was able to deliver more net content to a community using fewer resources. These compound data sets also allowed us to create trace-driven simulations that provided realistic evaluations of our innovative community-based networks.

In a more expansive approach to data synthesis, we present a methodological framework that integrates various modes of data collection and analysis to present a trace analysis nuanced by community knowledge. *Community data curation* is an iterative and collaborative process we used to collect traces that followed the connective actions of Native American advocates on Twitter. By partnering with community experts/gatekeepers, we

curated a list of entrypoints into the Twitter network, where entrypoints are hashtags, usernames or keywords suggested by community experts and experts they sampled (i.e., collected using a snowball sampling of a community of experts). Based on these initial entrypoints, we collected a cursory sample of Twitter content (tweet objects) associated with these entrypoints. Using data mining techniques, we identified salient keywords, hashtags, and users and we added these to our existing entrypoints. Using qualitative coding techniques, we identified themes, which we bring to community partners and academic experts to further unpack and add to our keyword entrypoints. We then used this much larger set of entrypoints to collect data generated by the Native American advocates over a six week period. Our analysis of the data we collected allowed us to present a semantically rich quantitative analysis of how marginalized groups appropriate social media platforms to form enduring connections via identity-based connective actions [204].

Indeed, our work makes specific contributions to the field in the form of specific findings and methodologies. Our work makes a larger contribution in its acknowledgement of the need for integrative, mixed methods approaches to understanding community interactions with information networks. Our creation of and reliance on these methodologies points to the fact that community members alone can provide necessary insight into their interpretations and uses of specific mechanisms of communication (such as tags, relationships, and media types) and it is only in partnership with communities that we as network scientists can accurately characterize how a community interacts with a network.

1.2.2 Network System Design

This dissertation presents network innovations in the forms of network infrastructure and mobile systems applications. Our work contributes to the state of the art in the

areas of computer supported collaborative work and pervasive computing systems, both with respect to our development of community-directed system resource allocation and systems that help users, designers, and organizations navigate the reality of heterogeneous network capacity and characteristics. We designed and evaluated an FM radio-based content delivery system. In this work, we created a metric of network structure that incorporates the graph concepts of in-degree and clustering coefficient to identify nodes with the widest and most widespread audience in a graph: the cumulative clustering coefficient. Specifically, we introduced an algorithm that prioritizes content based on the cumulative clustering coefficient associated with the creators of content. Using six months of Instagram traces generated by TDV users, we evaluated the fairness and coverage provided by the system. Our scheduling algorithm was able to provide half of the users in the community with 81% of their Instagram content requests and 35.5% of the 1.1 million requested Instagram photos were delivered to users over the six month simulation period. The key finding of this work is that community social networks can be leveraged to select relevant content for distribution over extremely low bandwidth.

We extend the concept community-driven resource allocation in our design of FiDO [203]. By combining the needs of rural tribal communities outlined in Chapter 2 with census and transportation data on the daily commuting habits of rural residents in the U.S., we design a content delivery system that augments cellular infrastructure with the capacity to fetch and store content on behalf of users from disconnected households so that they can opportunistically download that content on behalf of themselves and members of their households. This design is novel in that it integrates concepts of collaborative and content-based filtering from recommender systems to maximize the value of opportunistic cellular connectivity that mobile users encounter during a daily commute. In addition, we contribute the design of an innovative trace-driven evaluation approach, wherein we integrate network trace data with statistical models parameterized with rural mobility

data to provide a semantically realistic evaluation of FiDO's operation in rural Indian Country.

We also contribute to the design of two applications that seek to assist in the usage and design of mobile content creation and data collection applications in environments where connectivity to the Internet is poor or non-existent. First, we developed Submit as an extension of the Open Development Kit (a suite of applications for mobile data collection). Submit functions as a layer between mobile applications and a device's network interfaces. It provides abstractions to application builders, organizations, and users that assist in the selection of and transmission over diverse of available network interfaces. The selection of network interfaces depends on data properties, current network conditions, and organizational resource preferences. Concepts from the Submit platform were integrated into the Open Development Kit (ODK) 2.0 Tool Suite Aggregate Tables Extension to facilitate data collection in offline environments or in environments with limited Internet connectivity [143]. This integration promises to have significant impact on global data collection efforts as ODK has been used in deployments of data collection initiatives in 42 countries and an international space station [144]. We also make an impact contribution with the extension of VillageShare [107] to support collaboration and curriculum sharing across multiple VillageShare servers, enabling communities to establish collaborative content creation initiatives in schools and libraries [201]. Moreover, we develop a mobile application that enables users to create content and work on collaborative projects while disconnected from the Internet. The software for VillageShare servers as well as the software for the VillageShare mobile application are publicly available on GitHub at `https://github.com/VillageShare`.

Chapter 2

Research Background

Information is the most valuable currency of our time and significant power is derived from the ability to create, collect, own, and analyze information [34]. This information economy has given rise to a growing demand for information infrastructure. In 2016, the International Telecommunications Union (ITU) estimated that 47.1% of the global population used the Internet, but only one out of seven people in the least developed countries had access [99]. The majority of new subscriptions (particularly in developing countries) are mobile-broadband subscriptions, which have reached a penetration rate of 41% across developing countries. In addition to broadband becoming more widely available, it is also becoming more affordable; by the end of 2015, the price of 1 GB of computer-based mobile-broadband services was less than 10% of the *per capita* gross national index for developed countries and less than 20% of the *per capita* gross national index in the least developed countries. While this progress towards increased broadband access is encouraging, it fails to describe the variance of quality of access experienced by people connecting around the globe. According to the ITU, extreme variance in broadband speeds persists to the point where only half of providers in developing countries advertise data rates of 10 Mbps and fewer than 7% of providers in the least developed

countries advertise data rates of at least 10 Mbps [99]. This is particularly problematic as the sizes of content continue to grow due to a larger volume of dynamic Web content and higher fidelity screens, cameras, and other devices. Indeed, in the past six years, the size of the average Web page has grown from 760 KB to 2.5 MB [91].

Beyond the digital divide between developed and developing countries, there is a persistent divide that exists globally: the divide between rural and urban communities. In a 2015 study by the ITU, 89% of the world's urban population (4 billion individuals) had access to 3G or better mobile-broadband service while only 29% of the world's rural population (3.4 billion individuals) had access to the same level of service [98]. The lack of broadband infrastructure in rural areas has been addressed in numerous studies, which cite factors such as the low economic potential of sparsely populated areas, the expense of building infrastructure in remote and rugged terrain, the lack of supporting infrastructures such as a stable electrical grid, and the lack of maintenance personnel [21, 192]. Indeed, there is a pernicious digital divide in the U.S.; the Federal Communications Commission (FCC) estimates that only 15% of people living in Indian Country have access to broadband, with reports of coverage being much lower [142]. In contrast, the Internet access rate for the general U.S. population is 88.5% [100]. The digital divide in Indian Country sets the backdrop for this dissertation, which seeks to address the following question: *how can computer scientists improve information access in Indian Country?* To begin addressing this question, we provide some background information information about Indian Country in the U.S., tribal research partners, and our work's perspective on networks and communities.

2.1 Contextualizing Indian Country

"Indian Country is a legal term that refers to the federally-recognized tribes, state-recognized tribes, pueblos, rancherias, bands, and Alaska Native villages and corporations within the political boundaries of the U.S. Used colloquially and not in a legal sense whatsoever, Indian Country also refers to Native peoples habits and norms in this somewhat parallel society [58]." The lands that comprise Indian Country are mapped in Figure 2.1. Indian Country, in the legal sense, is the result of treaties made between the U.S. government and American Indian tribes. A tribe refers to any extant or historical clan, tribe, band, nation, or other group or community of Indigenous peoples in the U.S. Critical to the narrative of this research is the fact that most of these land treaties were made under duress, with tribes being forcibly removed from their homelands and relocated to less-desirable lands to make way for colonial settlers and government and corporate enterprises [60]. There are currently 567 federally recognized tribes and 326 distinct tribal lands recognized by the Bureau of Indian Affairs (BIA) [27]. In addition, as of 2008 there are 62 state recognized tribes and many more that are petitioning for state or federal recognition [112]. It also bears noting that not all tribes that are state recognized are federally recognized.

In 2010, the FCC initiated the National Broadband Plan with the goal of connecting 100 million disconnected Americans to broadband with 10 Mbps download speeds by 2020. As part of this plan, the FCC made specific recommendations for expanding access to broadband in rural areas including improvements on the Universal Service Fund (USF), creation of the Mobility Fund, and creation of the Connect America Fund (CAF)–all with input from tribal governments for issues pertaining specifically to broadband accessibility in Indian Country [65]. As we will discuss in the next sections, understanding the resulting governance structures, economic environments, infrastructural developments, and

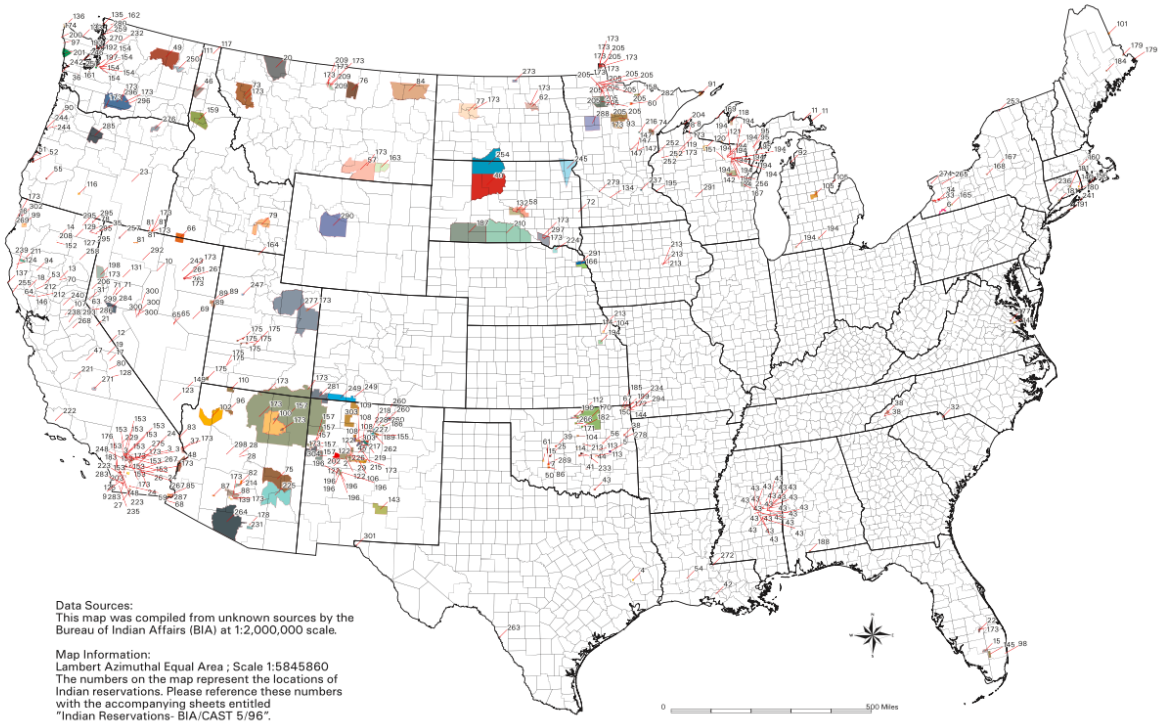


Figure 2.1: A map of the land in the U.S. that comprises Indian Country [137]. Colored areas represent reservations, trust lands, and rancherias recognized by the BIA.

cultural needs are critical to designing information infrastructure for Indian Country.

2.1.1 Tribal Sovereignty and Self-governance

One of the most significant consequences of federal recognition for tribes is tribal sovereignty. Tribal sovereignty allows for tribes to be recognized as self-governing dependent domestic nations located within the U.S.; as such, the U.S. federal government interacts with tribal governments on a government-to-government basis. Since the Indian Education and Self-determination Act of 1975, federal-recognition has also included self-determination, which enables tribes to create their own government programs and charges the federal government with assisting and supporting these efforts. In addition, the Act precludes federal intervention without prior consultation with and permission from the tribe. Sovereignty becomes more complex for state recognized tribes without

federal recognition. While state governments will grant these tribes self-determination and autonomy within the state, they still fall under federal jurisdiction and are not viewed as separate nations by the U.S. federal government. Critically, there is no federal obligation to uphold or support tribal programs for state recognized tribes.

For the deployment of telecommunications infrastructure, tribal sovereignty poses a unique challenge for telecommunications companies. Before they can begin installing any equipment or towers, companies must consult with tribal governments and perform environmental protection studies as well as historic preservation studies before finally consulting with the BIA [189]. This long, tedious process often acts as a deterrent for most telecommunications companies.

2.1.2 Infrastructure

Indian Country continues to be some of the least developed land in the United States. Two main reasons for this is the terrain upon which most reservations are situated and low population densities. According to Matthew Rantanen, a member of the FCC Native Nations Broadband Task Force and the Director of Technology for the Southern California Tribal Chairmen's Association (SCTCA),"When they made the reservation system in the federal government, they decided to put Indians where they thought nobody would want to be. They sent them to locations at the base of mountains or out in remote areas where the non-tribal population centers were, and they're far away from communication centers, where it's not advantageous to deploy infrastructure [189]."

Indeed, the cost of installing a mile of fiber cable for communications infrastructure costs $10,000 to $50,000 per mile [189] and many reservations have population densities of less than 50 people per square mile[1] [139]. Few telecommunications companies perceive deploying this type of costly infrastructure to serve only a few households as a cost-

[1]For non-tribal U.S., the average population density is 345 people per square mile [139].

effective investment.

Another result of low population densities, remote locale, and rugged terrain is a lack of existing utility infrastructure; most critically, electrical infrastructure. The Energy Information Administration estimates that 14% of households on tribal lands do not have electricity[2], a rate 10× higher than the average U.S. household [62]. This lack of infrastructure has several effects on initiatives to increase broadband penetration in Indian Country. First, it compounds the cost of deploying grid-dependent telecommunications infrastructure. Second, it often means that potential consumers may not have a traditional or reliable means for powering network devices or recharging mobile devices. Third, it means that many communities prioritize resources for establishing more basic utilities such as plumbing, sewage, and electricity before broadband infrastructure.

Finally, lack of a technically skilled workforce poses another infrastructural challenge to the deployment of broadband infrastructure in Indian Country. In 2010, only 8.4% of residents in Indian Country had obtained a college degree [4]. IT personnel capable of tasks such as network administration, technical support, systems analysis, and software development are necessary for the sustainable operation of broadband infrastructure; these tasks also require specialized skills that are typically only available via some post-secondary education and training.

One of the ways Indian Country has overcome the burden of limited infrastructure is through "leapfrogging", or skipping phases of infrastructure capacity-building in order to advance technologically. For example, reservation households might leapfrog by accessing the Internet at home through a mobile broadband network while still lacking access to a land line telephone connection. Leapfrogging has subtle implications for the quality of services available via infrastructure. As an example, 911 emergency services are dispatched with different speeds and accuracy when emergency calls are made from

[2]This percentage can be as high as 40% on some of the larger reservations [30].

land lines vs. cellular phones. In a different scenario, a home may have broadband access via a cellular network, but they the home may not have access to electricity. Although it presents some design challenges, leapfrogging also introduces unique opportunities. It allows communities to develop sustainable or nontraditional infrastructures, for instance, solar-powered microgrids [175] or modular cellular base stations [215, 90].

2.1.3 Economic Environment

Indian Country experiences some of the highest poverty rates in the U.S. Based on data from the 2010 Census, the real median household income[3] across all reservations was $30,023, compared to $51,076 for the total U.S. population. The family poverty rate for all reservations was 33.5%, compared to 10.1% for the total U.S. population. Additionally, the unemployment rate on Indian reservations is 12 percentage points higher than for the total U.S. population [4].

This poverty gap has significant implications for the state of telecommunications infrastructure. First and foremost, monthly broadband fees can be a burden for many families and those in Indian Country interested in paying for broadband services may not have the financial wherewithal to do so consistently [139]. In addition to some of the infrastructural barriers mentioned in Section 2.1.2, this poverty gap discourages many commercial broadband providers from investing in infrastructure based in Indian Country. Moreover, much of the land in Indian Country is not individually owned (i.e., it is held in trust by the tribe), meaning that it cannot be used to establish credit for loans that could enable communities to develop infrastructure on this land [68].

[3]Real income refers to the income of an individual or group after taking into consideration the effects of inflation on purchasing power.

2.1.4 Cultural and Political Significance of ICTs

A discussion on ICTs in Indian Country is incomplete without a brief overview of some of the unique cultural perspectives on governance, infrastructure, and communication. Traditional cultural knowledge has suffered significantly in the centuries-long conflict between Native Americans and the U.S. federal government [60], including the disappearance of native languages, eradication of religious practices and traditions, and loss of family histories. Moreover, contemporary Native American culture is rife with memories of its recent oppression via massacres, legislation, forced removal, and Indian boarding schools [178]. This has resulted in historic trauma experienced on a mass scale, manifested in the high rates of suicide committed by Native American youth, violent deaths of Native American youth, and the prevalence of diseases such as alcoholism and diabetes [60]. However, there is impressive resilience to acknowledge. Survey and interview-based studies of Native American ICT usage reveal that Native Americans utilize the Web, and social media platforms in particular, to create cultural identity, foster resilience, and to preserve traditional culture [129, 76, 31].

Indeed, in his work on identity and network society, Manuel Castells predicted that as ICTs became increasingly available, identity-based groups would utilize these this technology to organize politically and communicate their marginalization to a global audience [34]. As the prevalence of ICTs has increased over the past three decades, this has proven true [129]. One of the most recent examples of this is how in 2016 and 2017, Standing Rock protesters instigated a global movement to protect the land and water rights of the Lakota people in the face of governmental and corporate mining interests [168]. Yet, as Duarte asserts in a study of Indigeneity in a network society, it is important to remember that Native American political discourse (especially as it pertains to self-governance and land rights) are not identity-based movements, but represent a

long-term flexible mode of governance for a land-based people [58]. Thus, understanding and designing ICTs for tribal communities in the U.S. requires that we acknowledge the core importance of a people's hard-fought self-governance over lands to which they belong in a way that is culturally foreign to colonially established governments. As observed by a Supreme Court justice in 1960:

> It may be hard for us to understand why these Indians cling so tenaciously to their lands and traditional tribal way of life. The record does not leave the impression that the lands are the most fertile, the landscape the most beautiful or their homes the most splendid specimens of architecture. But this is their home-their ancestral home. There they, their children and their forebears were born. They, too, have their memories and their loves. Some things are worth more than money and the costs of a new enterprise [141].

Native American peoples will not abandon tribal lands because it lacks infrastructure. Instead, there is a willingness to build and create infrastructure that is owned and controlled by the tribe in accordance with the high value placed on self-governance. The challenge for technologists lies in coming alongside these communities and understanding how they can participate in the creation of technological designs that allow tribal communities to build their own information systems, policies, and programs to meet tribal self-governance goals.

2.2 Tribal Partnerships

In order to begin understanding the information and technology needs in Indian Country, we formed partnerships with two different Internet Service Providers (ISPs) that provide broadband to homes, municipal buildings, and businesses on reservations.

2.2.1 Tribal Digital Village Network

In the 1990's, the Southern California Tribal Chairmen's Association (SCTCA) began ideating solutions to address the lack of Internet access to its tribal communities. The resulting Tribal Digital Village (TDV) network was conceived in partnership with the High Performance Wireless Research and Education Network (HPWREN), a broadband Internet network designated for scientific use by scientists at the University of California in San Diego, and Hewlett Packard labs. With a three-year five million dollar grant from Hewlett Packard and the assistance from engineering consultants from Hewlett Packard and HPWREN, the TDV program was established in 2001 [58, 166]. In the subsequent sixteen years, the TDV program deployed 350 miles of point-to-point and point-to-multipoint wireless links that provide Internet access to tribal homes and municipal buildings located throughout 17 different reservations. There are 2,700 homes on the reservations and about 2,500 more in the surrounding non-tribal areas. Currently, about 10% of homes on the reservation are connected to the Internet via the TDV network. One of the critical functions of the TDV network is to provide a competitively priced provider option for reservation residents and programs. There are two tiers of service available to residential customers: $34.95 per month for 2 Mbps or $64.95 per month for 3 Mbps; community anchor institutions such as schools, libraries, and tribal offices receive service for free [179, 166]. We established a research partnership with the SCTCA in 2014 through Matthew Rantanen, the Director of Technology to the SCTCA. Rantanen and his team of network administrators, including Geoff Herrin and Joseph Peralta, have served as collaborators in a significant portion of this work.

Geographically, the TDV network is located in eastern San Diego County in California. The TDV network is entirely wireless; a 500 Mbps fiber link connects to TDV headquarters in Pala, CA and is extended to other reservations over multiple 11 and

18 GHz microwave links. From the towers in the backbone, connectivity is extended to access towers located on reservations using 2.4 and 5 GHz . Finally, connectivity is extended to homes, businesses, community buildings, and tribal offices via 2.4 and 5 GHz WiFi.

2.2.2 Red Spectrum Communications Network

Red Spectrum Communications was established to serve the 3,500 residents of the Coeur d'Alene Indian Reservation. Created out of a joint venture by U.S. Army veteran and IT specialist Valerie Fast Horse and IT specialist Tom Jones, Red Spectrum Communications began as a Wireless Internet Service Provider (WISP) in 2002 using funding from the Coeur d'Alene Indian Tribe and the U.S. Department of Agriculture (USDA) Rural Utility Services (RUS) Community Connect grant and loan program. As part of their ongoing survey and analysis of broadband needs in the Coeur d'Alene tribal community, Fast Horse and Jones also expanded Coeur dAlene IT services to include a geographic information services (GIS) program for surveying and managing tribal lands, waters, and broadband access for the Coeur d'Alene tribe; their GIS database is one of the most thorough with regards to broadband connectivity status in Indian Country. In 2009, given their ability to demonstrate increasing demand for high-speed Internet and increased broadband penetration, Fast Horse and Jones received funding from the tribal council and the USDA American Recovery and Reinvestment Act (ARRA) grant and loan program to deploy 275 miles of terrestrial fiber-optic cable to supply affordable broadband Internet to 3,500 households within the Coeur d'Alene reservation and neighboring communities [58]. At the time of this work, the Red Spectrum Communications network provides service to 1,011 subscribers, 95.7% of which are residential and 4.3% of are community anchor institutions (e.g., schools, health centers, libraries, tribal offices).

Red Spectrum Communications provides four tiers of service: $34.95 per month for 3 Mbps, $59.95 per month for 10 Mbps, $79.95 per month for 25 Mbps[4], and $124.95 per month for 100 Mbps[5] [162]. Our research partnership with Red Spectrum Communications began in early 2016; Fast Horse, Jones, and network manager Justin Hall have all been collaborators in our analysis of the Red Spectrum Communications network, which we shall refer to hereafter as the Red Spectrum network.

Geographically, the Red Spectrum network is located in western Idaho and eastern Washington state. It provides Internet services to the Coeur d'Alene Indian Reservation, which is comprised of several small towns with population sizes that range from 175 to 1,026 [139]. In addition to providing broadband to towns located on the Coeur d'Alene Indian Reservation, Red Spectrum provides broadband to the towns of Mica, WA and Fairfield, WA, which are not located on tribal land and have a significantly different demographic composition than the towns located on the Coeur d'Alene Indian Reservation.

2.3 Communities and Networks

In the previous section, we discussed communities and networks as they relate to Indian Country. In this section, we define and describe communities and networks as they generally pertain to the analysis and design of information systems in challenged environments.

[4]Available to fiber customers only.
[5]Available to fiber customers only.

2.3.1 Communities

In its most abstract form, a community is a group of people with something in common. Just a few examples of this point of commonality include: location, citizenship, religion, language, culture, interests, philosophy, or genetics. In our analysis of how computer networks are utilized (see Section 2.3.3), we characterize usage through the lenses of social community, geographic community, and interest-based community. When examining usage using social community, we rely on online social networking interactions between individuals to define the bounds of communities; when characterizing geographic communities, we rely on geopolitical boundaries to identify communities; and when examining interest-based communities, we use online social networking tagging and content propagation to identify topical communities.

2.3.2 Networks as Data Structures for Relational Information

One way of understanding a network is as a data structure for storing information about the relationship between different objects[6]. Objects are represented as nodes in the network and edges represent relationships between nodes. Edges take on a number of properties that impact the interpretation of the network as a whole. For example, edges might be directed or undirected and edges might take on different values that indicate the weight of the relationship or the distance between nodes. It is also noteworthy that networks with similar node and edge semantics can be connected together to create even more complex networks.

This concept of networks is essential to our analysis and characterization of communities in Part I. We examine relationships between individuals, between individuals and institutions, between individuals and content, between individuals and geography,

[6]In computer science, this is also referred to as a graph.

and between individuals and community through the network model. In addition, we examine how some of these relationships alter over time using networks of networks.

2.3.3 Networks as Information Infrastructures

Computer networks are systems wherein the nodes are hosts and the edges are communication links which consist of various physical media, including optical fiber, copper wire, coaxial cable, and radio spectrum. Depending on the physical medium and its specific characteristics, data transmitted across a communication link can be sent at different rates, measured in bits per second. Computer networks enable a connected group of computers to share resources such as data, storage capacity, and computational power [114].

Computer networks are increasingly ubiquitous and pervasive. The most well-known computer network is the Internet, which connects hundreds of millions of computing devices around the world and an increasing number of devices, including mobile and wearable devices, require Internet connectivity for basic operation. Just as network data structures can be comprised of smaller network data structures, the Internet is comprised of smaller networks. Kurose and Ross identify a few types of networks that comprise the Internet: enterprise networks, national or global networks, local or regional networks, mobile networks, and home networks. Given our emphasis on challenged environments, we focus on local networks in the form of local Internet Service Providers (ISPs) and local broadcast radio stations, mobile networks, and home networks. Per the fact that many challenged environments rely on technology leapfrogging, we focus predominantly on wireless technologies. Our focus on wireless communication links informs the specific models we use to evaluate the design of innovative network systems.

2.4 Discussion and Conclusion

Indian Country is a unique environment that represents a microcosm of the issues experienced by many rural and developing countries around the world: economy, regulatory hurdles, lack of infrastructure, and cultural ideals surrounding ICTs [21]. Indeed, many "developed" countries with colonial histories have forms of this resource divide (e.g., the Aboriginal Australians, the Mixtec of Mexico, and the Sami of Norway). Based on historic patterns of global infrastructural development, it is realistic to assume that there will always be a lag between the initialization of new technologies and supporting infrastructures and the global adoption of these technologies. It is arguable that due to our increasingly networked global society [34], the consequences of that lag will be felt ever more acutely by those countries and communities that fall on the tail end of infrastructure development. Indian Country is a particularly interesting example of this effect in that it is located within the borders of a very wealthy and generally developed nation, implying that development measured on a national level does not sufficiently ensure rights to access for all and may mask severe inequities among divergent groups.

2.4.1 Conclusion

In this dissertation, we demonstrate that by partnering with ISPs operating in Indian Country, we as information system architects can get a sense for how broadband is used when it is available, the benefits of which are twofold. First, we are able establish a community-centric baseline for how people living in Indian Country *want* to use broadband. There is a long history of outside agencies incorrectly assuming this knowledge to the detriment of the self-governance goals of tribes and welfare of tribal people (e.g., the BIA's role in tribal resource management prior to the Indian Self-Determination and Education Act of 1975). There is also a history of marginalized people (Indigenous peoples

in particular) appropriating technologies for utilization that transcends their original intent [149, 6, 188, 44, 207, 110]. We must humbly participate with communities to learn their goals and preferred avenues of attainment before we critique (as academics) or offer solutions (as technologists). By combining our observations with an awareness of limitations as well as the optimizations possible through the utilization of network data structures and infrastructures, we can help design information systems that are well-tailored for Indian Country.

Part I

Characterizing Usage with a Network Analytic Approach

Chapter 3

Web Usage in the Tribal Digital Village and Red Spectrum Networks

In Chapter 2, we outlined how broadband played a role in economic development, access to education, and strengthening cultural identities for marginalized communities. One of the most common applications of broadband is the Web. Since its inception in 1989, the Web has fundamentally changed how people interact with each other, share information, and entertain themselves [156]. For Indigenous communities, the Web (when accessible) has enabled transnational organization [6, 207, 55, 125, 60]. In this chapter, we present the first observational and quantitative studies of Web usage in an Indigenous context.

3.1 Data Collection

For both the TDV and Red Spectrum networks, Web data is mined from traces collected at the Internet gateway of the network. We collect traces by attaching a traffic monitoring server to the switch that bridges the gateway and the network. A mirror port on the switch is configured to capture all packets traversing the gateway link. We

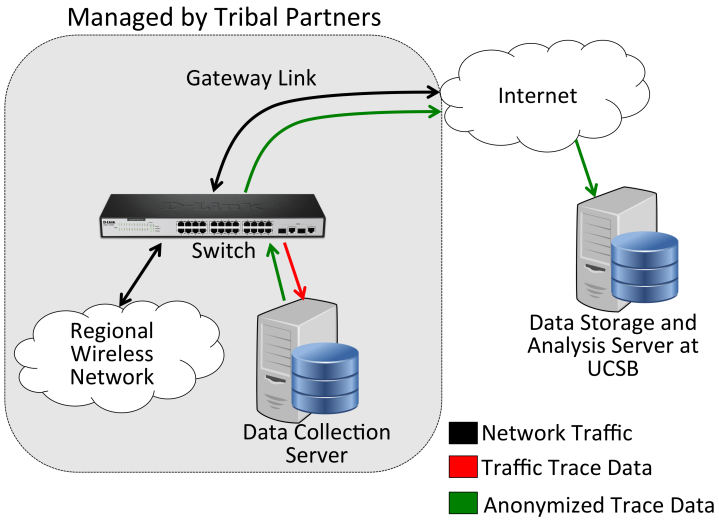


Figure 3.1: A diagram depicting the various components of our Web data collection methodology.

capture packet headers with `tcpdump` [103] and use the Bro Network Security Monitor to collect flow-level statistics for network applications and protocols [23]. All MAC addresses and IP addresses are anonymized using the prefix-preserving anonymization tool, TraceAnon [184]. After the network data traces have been collected (depicted as the red arrow in Figure 3.1 and anonymized at the data collection server, we transfer the anonymized traces for long-term storage and analysis on a server at UC Santa Barbara (UCSB) (depicted as the green arrows in Figure 3.1). We ensure that data transfer does not effect the performance of our partners' networks by capping our data transfer rate to 2 Mbps. Network traces traversing the gateway link are not included in subsequent analysis and reports of traffic in the TDV and Red Spectrum networks. An illustration of our data collection process is shown in Figure 3.1.

We present a timeline of our data collection in Figure 3.2. Over our 3 years of partnership with the SCTCA, we have collected 114.8 TB of data logs from the TDV network. Our work characterizing Web traffic examines three representative samples of this larger data set, which we will refer to as `TDV Web-2014`. In our four months of

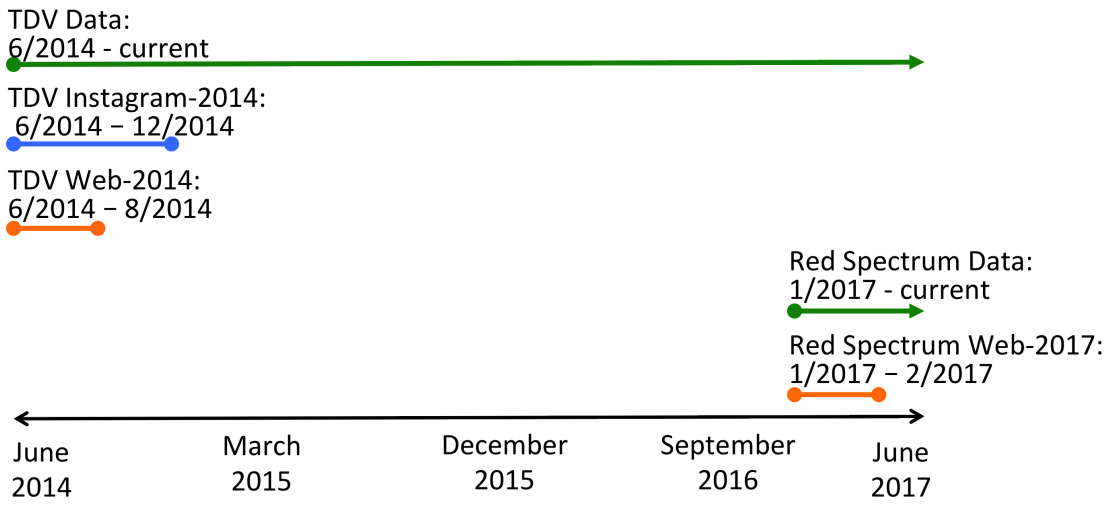


Figure 3.2: A timeline of our data collection and temporal contextualization of the analyzed subsets.

Table 3.1: Overview of the traffic profile associated with each of the data sets used in our analysis.

	Date range	Total traffic volume	% Web volume	# packets	# Web transactions
`TDV Web-2014`	6/23/14 to 8/20/14	29.52 TB	87.5	31.3 billion	23.9 million
`Red Spectrum Web-2017`	1/17/17 to 2/28/17	23.9 TB	95.7	24.6 billion	8.3 million

partnership with the Red Spectrum Communications, we have collected 728 GB of data logs from the Red Spectrum network. Our work characterizing Web traffic examines a representative sample of this larger data set, which we will refer to as `Red Spectrum Web-2017`. We provide an overview of the data sets we collected in Table 3.1.

Table 3.2: Devices used in TDV network.

Mobile devices		411
	iOS	42.9%
	Android	35.2%
	Windows	11.5%
Desktop devices		131
	Mac OS X	23.1%
	Linux	11.9%
	Windows	45.0%

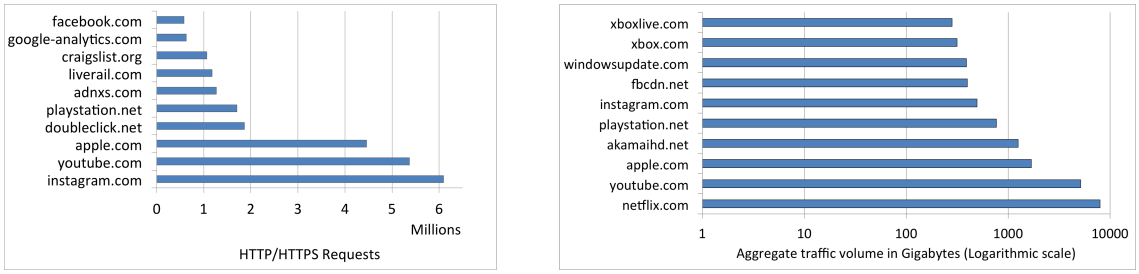
Table 3.3: TCP statistics for each reservation.

	Link	% Failed requests	% Retransmitted	RTT (ms)	IAT (s)
Pala	–	25.6	5.9	140	0.37
Pauma	L1	24.0	5.7	89	0.26
Rincon	L1	28.8	4.6	93	0.26
San Pasqual	L1	33.9	4.9	94	0.30
Mesa Grande	L2	25.7	4.8	71	0.15
Manzanita	L5	23.1	5.1	79	0.27

3.2 Analysis of the `TDV Web-2014` Data Set

Table 3.2 provides an overview of unique devices used to access the Internet over the TDV network. These devices were identified using the user agent field of HTTP traffic headers. The majority of users access Web content using mobile devices (smart phones, tablets, e-readers) as opposed to stationary devices (desktops, laptops, and gaming consoles). Additionally, gaming consoles account for 11% of desktop devices in the TDV network.

In order to characterize traffic, we examine network performance both as a whole and associated with each reservation. In Table 3.3, we report performance statistics associated with each reservation relay link. The 'Link' column in the table is associated with



(a) (b)

Figure 3.3: Top 10 Web domains in the `TDV Web-2014` data set based on (a) the total number of HTTP/S requests and (b) the aggregate traffic volume.

the backbone link that connects to each relay tower. Retransmission rate is calculated according to the number of retransmitted segments in a flow divided by the total number of segments transmitted. We observe high retransmission and failure rates at all relay links. Pala, which connects to the gateway via a single relay link, has the highest failure and retransmission rates. Exemplified by Pauma, Rincon, and San Pasqual, performance can vary in terms of packet loss and flow failure even for reservations connecting to the backbone via the same link. Likewise, reservations that are multiple hops away from the gateway, such as Mesa Grande and Manzanita, do not experience performance degradation in proportion to the number of backbone hops they must travel. Based on these observations, we conclude that performance degradation occurs either over the relay links between backbone towers and individual reservations or over the access links that extend connectivity from relay links into homes and municipality buildings. Our current measurement configuration does not allow us to pinpoint the exact location of packet loss, but based on our findings and our knowledge of the network topology, we suspect that packet loss is either due to low signal strength between the backbone towers and relay points or due to interference between access links connecting to the same relay point.

3.2.1 Applications and sites accessed

Figure 3.3a reveals that Instagram is the most requested Web application, with over 6 million requests over the measurement period. This is surprising given current measurements of Web usage in the U.S., which have revealed Facebook as the most dominant social media presence, followed by Twitter, LinkedIn, and Pinterest [5]. We are also surprised by the high levels of gaming traffic represented by PlayStation and Xbox. While Xbox is not ranked as one of the top 10 most requested applications in the network, it is ranked in the top 15. Similar to what we observed with the rank of Instagram traffic, the prevalence of PlayStation is unexpected; it ranks as the sixth most accessed Web site in the TDV network compared to 1,089th in the U.S. [5]. In terms of popularity, YouTube is second only to Instagram with 1.8 million requests during the measurement period. This is similar to the rest of the U.S. where YouTube is ranked as the third most popular Web site [5]. Surprisingly, e-commerce sites like Amazon and Ebay are not even ranked in the top 30 most accessed Websites in the TDV network, despite ranking as the 4th and 8th most accessed Websites in the U.S. at the time the `TDV Web-2014` data set was collected. When searching for an explanation for this dearth, we find that sovereignty plays a role. Many e-commerce sites rely on U.S. Postal Services for shipping; however, as reservation roads are not maintained by the county, U.S. Postal Services will not deliver to homes that must be accessed through these roads. Similarly, other major shipping companies (i.e., UPS and FedEx) reserve the right not to deliver to all areas and do not guarantee delivery to all addresses. In short, online ordering and shipping can be a significant challenge on reservations and the lack of e-commerce traffic in the TDV network reflects this.

Figure 3.3b shows the top 10 most bandwidth-consuming Web sites observed in the TDV network. We find that streaming media sites represent the greatest bandwidth

consumers in the network, accounting for 44% of the overall Web traffic volume. This is consistent with streaming media usage in the rest of the U.S., where streaming media accounts for 34-50% of peak traffic bandwidth [97]. Likewise, the composition of streaming media mirrors that of the U.S., where Netflix accounts for 60% of streaming media. We also notice that video gaming sites such as xboxlive.com rank among the top 10 Web sites in terms of bandwidth and overall, gaming traffic accounts for 8.9% of Web traffic volume. Social media comprises a much smaller portion of traffic volume, counting for only 4.5% of the total bandwidth consumed by Web traffic. Finally, media and software downloads from online stores such as the Google PlayStore and the iTunes Store account for 23% of Web traffic volume. This is in contrast to global mobile traffic patterns, which measure app store downloads as accounting for less than 18% of mobile traffic [97]. In total, media represents over 70% of observed Web traffic by volume. Given the popularity of media-oriented applications in terms of HTTP requests and proportion of traffic volume, as well as tribal interests in leveraging broadband for the development of cultural media, we focus the remainder of this paper on understanding media performance and usage in the TDV network.

3.2.2 Performance of select applications

Previous studies of emergent wireless networks have observed connections between usage patterns and network performance [106, 216, 39]. Based on the high levels of media traffic in our initial examination of the TDV network with the `TDV Web-2014` dataset, in addition to the popularity of media-based sites, we look specifically at the performance of media applications in the TDV network to identify the presence of performance bottlenecks that might impact user behavior. In order to do this, we look to the performance of three applications: YouTube, Netflix, and Instagram. These three

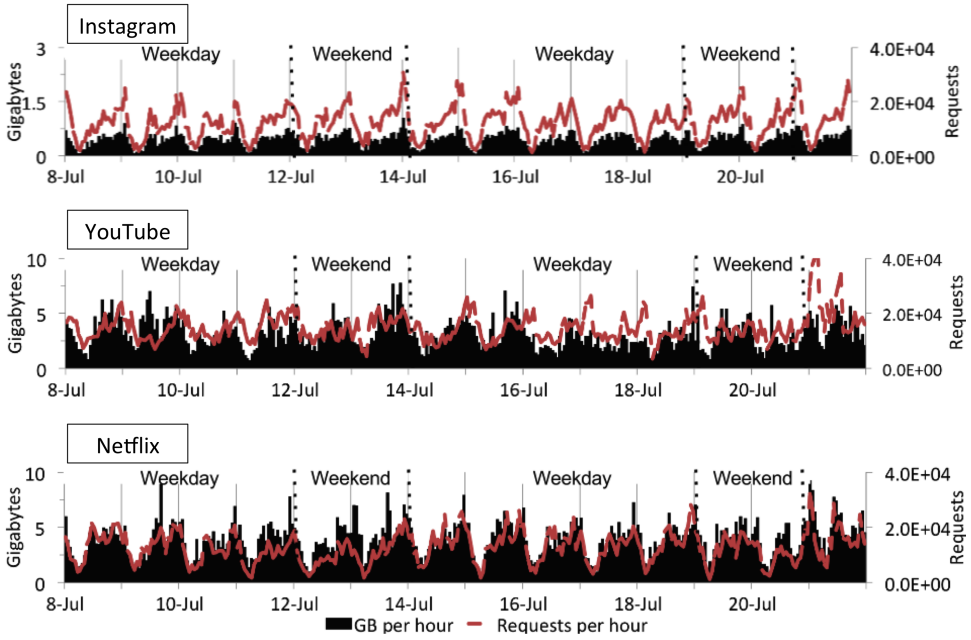


Figure 3.4: Hourly traffic demand for Instagram, YouTube, and Netflix from July 8 to July 21.

applications are representative of the predominant media transaction types (streaming and bulk transfer).

We select each application based on its high traffic volume in the network and the different ways that it allow users to interact with media. We begin by studying the daily usage of each application. In Figure 3.4, we show the daily traffic volume and Web requests generated by Instagram, Netflix, and YouTube over a two-week period which we verify as representative of the entire sampling period (note the different y-axis scales). Traffic volume per hour was calculated by summing the total number of bytes per hour; the number of HTTP requests per hour was calculated by summing the total number of HTTP requests per hour. While weekend and weekday traffic are not significantly different for any of the applications, all three applications exhibit anthropocentric patterns in their usage over time.

Table 3.4: TCP statistics for Instagram photos, Instagram videos, YouTube videos, and Netflix videos.

	Retransmitted %	Failed downloads %	Failed uploads %	RTT (ms)	IAT (s)
Instagram photos	3.2	11.6	24.9	91	0.22
Instagram videos	3.8	31.0	25.0	91	0.29
YouTube videos	2.9	32.3	30.1	73	0.18
Netflix videos	3.4	75.0	NA	94	0.34

We now look at media performance for Instagram, YouTube, and Netflix. Table 3.4 reports performance statistics for each application including the retransmission rate, download failure rate, upload failure rate (when applicable), round trip time (RTT), and packet inter-arrival time (IAT). Round trip times and packet inter-arrival times associated with each application were calculated by taking the average round trip time and packet inter-arrival time for each TCP flow and averaging them over all TCP flows. The retransmission rate for each application represents the percentage of segments retransmitted per TCP flow averaged across all TCP flows.

Video Downloads. Understanding the user experience with regards to the network performance of these platforms can highlight areas of improvement in terms of network infrastructure and application design. Overall, we find relatively high rates of download failure for YouTube, Netflix, and Instagram videos. We find that Netflix has the highest failure rate of all, with 75.0% of video downloads ending in a failure. In comparison, 32.3% of YouTube downloads fail, and Instagram video downloads experience a failure rate of 31%. In exploring failures associated with video downloads, we find that the predominant cause of failure for all applications is a TCP RST sent by the client. This type of failure causes 63% of download failures for YouTube, 60% of download fail-

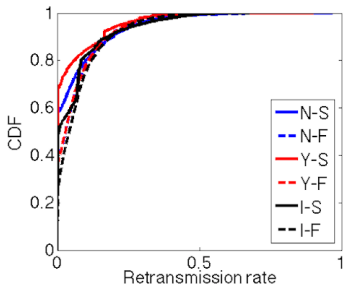
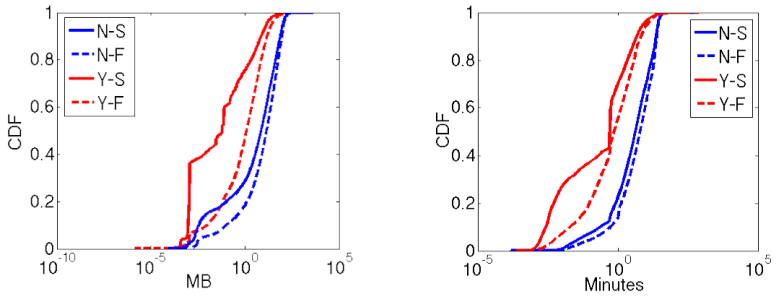


Figure 3.5: Cumulative distribution of retransmission rates for downloads.



(a) (b)

Figure 3.6: Cumulative distributions of the (a) flow sizes and (b) durations of Netflix and YouTube downloads.

ures for Netflix, and 58% of download failures for Instagram videos. This type of failure is indicative of a poor user experience, often due to packet loss [106].

To assess the impact of retransmission rate on download performance, we compare the distributions of successful video downloads to failed video downloads for Netflix (N), YouTube (Y), and Instagram (I) in Figure 3.5. As an example of our notation, we use "N-S" to signify successful Netflix flows and "N-F" for failed Netflix flows. We find that for all three applications, retransmission rates are higher for failed downloads than for successful downloads, and on average, failed flows experience 8.2-9.7% loss. For streaming video applications, such as YouTube and Netflix, this type of loss rate would negatively impact the user quality of experience and termination of download mid-stream

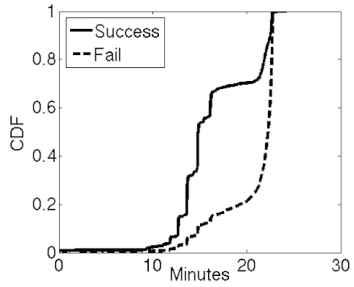


Figure 3.7: Cumulative distribution of flow durations for Instagram video downloads.

(triggering a client-sent RST) is consistent with our findings of cause of failure. To assess how failure impacts user interaction with streaming video applications, we study the distributions of flow size and flow duration for successful and unsuccessful Netflix and YouTube downloads using Tstat [158]. Figure **??**a graphs the distribution of flow sizes for failed and successful flows for both YouTube and Netflix and Figure 3.6b shows the distribution of flow duration for failed and successful flows for the two applications. For the size of a flow, we report the goodput rather than the total size including retransmitted data. We find that for both applications, failed flows are on average 83% larger and last 28% longer than their successful counterparts. While we have established that failed downloads are associated with higher retransmission rates, we note that these failures correspond to longer, lengthier downloads, which are more likely to experience losses resulting in a poor user experience. However, we are surprised to find that although Instagram videos are smaller than YouTube and Netfix downloads and experience shorter flow duration, this application experiences the highest retransmission rate for both failed and successful downloads. One reason for the high retransmission rate for successful Instagram video downloads is that it downloads media in bulk, rather than as a stream– so quality of experience is not impacted by the number of packet losses. However, packet loss impacts the length of time it takes to download a video before a user can watch it. In Figure 3.7, we compare the distributions of flow duration for successful and failed

Instagram video downloads. We find that download times are 62% longer for failed downloads than for successful downloads. This is expected as packet loss leads to more retransmissions that result in longer download times. Overall, 40% of downloads require over 15 minutes.

Video Uploads. We now investigate upload performance for YouTube and Instagram videos. Upload performance in these applications is particularly important in the context of the TDV network given many of the goals tribal communities have for broadband connectivity, including cultural content creation, dissemination, and engagement. Overall, 504 video files were uploaded to Instagram compared to 444 uploaded to YouTube. 25% of Instagram video uploads failed and 30% of YouTube uploads failed. When examining the predominant cause of failure for uploads, we find that for Instagram, 85% of failures were due to an unresponsive client (no data packets or control packets were observed coming from the client) and 11% of failures were caused by an RST sent by the server. For YouTube, 55% of upload failures are caused by a timeout-triggered RST sent by the server and 36% of failures are caused by an unresponsive client.

While packet loss over relay and access links can be a contributing factor to failure, we also consider that all Instagram uploads and 98% of YouTube uploads are initiated from mobile devices. Uploading from a mobile device increases the likelihood of a user inadvertently moving from a space of high connection quality to low connection quality, particularly if upload times are extensive. On average, successful Instagram video uploads take 1.4 minutes and successful YouTube video uploads take 4.9 minutes. Failed Instagram uploads take 2.1 minutes before termination and failed YouTube uploads take 5.7 minutes before termination. We believe that this short upload duration, facilitated by lower retransmission rates for upstream traffic, is what allows for Instagram video uploads to be more successful than video downloads. The average uploaded Instagram video is only 0.24 MB compared to the average uploaded YouTube video, which is 11.2 MB.

This difference is unsurprising given the restrictions Instagram places on video length (15 seconds), dimension (640×640 pixels), and resolution. This is in contrast to YouTube, which limits video uploads at 11 hours or 128 GB. Increased video sizes have negative consequences for upload failure, as they typically take longer to upload and are more likely to experience packet loss. We also find that on average, failed Instagram uploads are retried 1.3 times and failed YouTube uploads are retried 1.7 times. 3% of retried Instagram video uploads and 7% of retried YouTube uploads are never successfully completed. For each of these failed retries, we find that each final attempted retry takes 10 minutes on average before the attempt is terminated. We also find that repeatedly failed Instagram video uploads require 3 more minutes than video uploads that eventually succeed; repeatedly failed YouTube uploads require 3.6 more minutes than video uploads that eventually succeed.

Images. With the largest number of uploaded files and a platform that lends itself to media sharing and collaboration, Instagram embodies much of the community enrichment potential of broadband connectivity. Unlike YouTube, it enables users to upload images as well as videos. Because the Instagram app was designed exclusively for mobile devices, it imposes limitations on its media formats and as a result provides a higher likelihood of successful media transfers. This primarily manifests as limited upload file dimensions (640×640 pixels). Once an image has been uploaded to Instagram, three standardized versions are created: small image ($\geq$10 KB), medium image ($\geq$20 KB), and large image ($\geq$100 KB). For uploaded videos, small, medium, and large images are generated from the first frame of the video. When the Instagram app is active, images are downloaded from content servers where large versions of the image correspond with images that are more relevant to users, medium images correspond with images that are less relevant, and small images are used for metadata reports. In Table 3.5, we show the failure rates associated with small, medium, and large downloaded, as well as uploaded

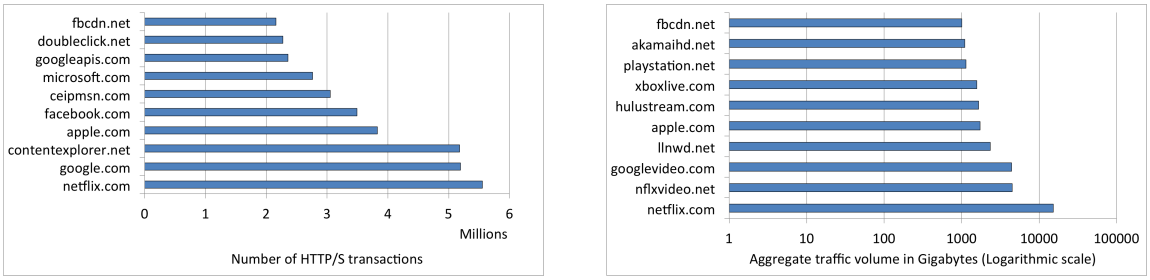
Table 3.5: Failure rates for Instagram images.

	Total #	Failure rate (%)
Small images	5,450,807	13.9
Medium images	2,880,253	15.0
Large images	2,345,020	17.6
Uploaded images	10,677	23.6

Instagram images. 76% of download failures and 86% of upload failures are caused by unresponsive clients after a TCP connection has been established.

3.3 Web Preference Similarities in the `Red Spectrum Web-2017` Data Set

Just as we did in Section 3.2.1, we examine the top ten most accessed Web domains and the top ten Web domains based on traffic volume in the Red Spectrum network in Figure 3.8. Similar to the TDV network, the most prevalent domains are associated with streaming media (netflix.com and hulustream.com), social media (facebook.com and fbcdn.com), and gaming (xboxlive.com and playstation.net). Also similar is the fact that e-commerce sites are not present in the top 20 most accessed domains in the `Red Spectrum Web-2017` data set despite the fact that amazon.com and ebay.com were ranked 5 and 10 respectively for the general U.S. population [5]. A difference is that facebook.com is the most accessed social media site in the Red Spectrum, as opposed to instagram.com in the TDV network. Further investigation is needed to discern whether this difference is due to regional differences in platform preference or if it is due to the fact that subscribers in the TDV network are more likely to have a lower data rate connection to the Internet than subscribers in the Red Spectrum network (see Section 2.2). One of the main purposes for our investigation into the Red Spectrum is to understand how



(a) (b)

Figure 3.8: Top 10 Web domains based on the (a) number of HTTP/S transactions and (b) total traffic volume.

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

$$C(A, B) = \frac{A \bigcap B}{A} \tag{3.1}$$

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 3.9a, 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.

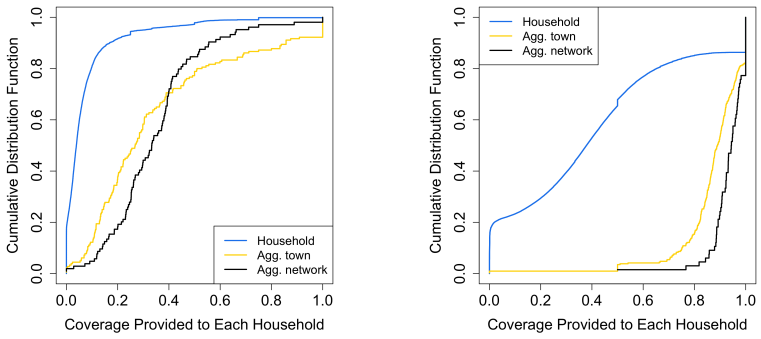


Figure 3.9: Cumulative distributions associated with (a) the file coverage and (b) the domain coverage provided by different scopes of community in the Red Spectrum network.

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.

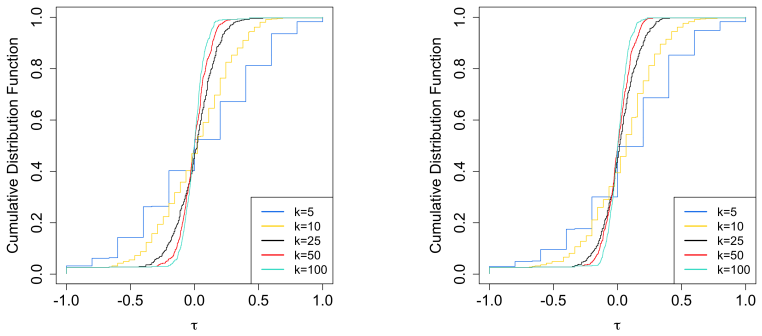
In addition to file coverage, we seek to characterize the community's ability to provide Web content for disconnected households by examining the *domain* coverage provided at different scopes. Instead of measuring coverage on specific file content, we measure it with respect to the Web domains visited 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 3.9b 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 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 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 [63], which is calculated by:

$$\tau = \frac{\#\ concordant\ pairs - \#\ disconcordant\ pairs}{k(k-1)/2} \tag{3.2}$$

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



(a) (b)

Figure 3.10: Cumulative distribution associated with Kendall's τ similarity between (a) the top k household domains and top k town domains and (b) the top k household domains and top k global domains in the Red Spectrum network.

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 3.10 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.

3.4 Discussion and Conclusion

In this chapter, we present our analysis of Web data collected from the TDV network and the Red Spectrum network. Here we reflect on our methodological process and results and discuss the most prominent points.

3.4.1 Challenges Associated with Web Traffic Analysis

As discussed in Chapter 2, the Web trace data we collect is generated by contextually specific communities. As we look towards extending these types of analysis methodologies, we observe three significant challenges. First, because we only collected header-level information, our analysis is limited to information and meta-data available in packet headers. This challenge is not prohibitive; for example HTTP packet headers contain information about host names, URLs, referers, user agents, and the MD5 hashes of files being transferred over HTTP. Even without the full packet, these header fields can provide a significant amount of information, such as which domains users visit, which applications are most utilized, and which devices are most common. The second challenge is that an increasingly large portion of Web transactions are conducted over SSL. This means that the majority of header fields are encrypted, including full URLs, user agents, and referers. This leaves us with a significantly diminished set of data to analyze. For example, there is a significant amount of meta-data present in full URLs. Information such as account identifiers and user identifiers can allow us to form edges between users

and locations in the network, users and specific Web domains, and between users themselves. The fact that the majority of traffic volume represented by the `TDV Web-2014` and `Red Spectrum Web-2017` data sets represents unencrypted data allows us to address this challenge. However, trends towards applications' increasing reliance on SSL suggests that we will no longer be able to perform the same type of analysis. Looking to the future, there will be a need for methodologies that can infer basic Web traffic profiles. The third and most significant challenge is the ever-changing nature of the Web. Applications are updated and meta-data that was once embedded in URLs, referers, or browsing agents, is no longer accessible in packet headers. This makes semantically rich analysis of Web packet headers a dynamic process that must constantly evolve to match the information available.

3.4.2 Need for Spectrum Access

While previously studied networks in developing contexts typically reside within greater regions of development, tribal communities in the U.S. are positioned within a developed nation. In many ways, the effects of the digital divide are amplified in this context and citizens without broadband access are marginalized at an accelerated rate. With high broadband penetration rates in the non-tribal U.S., many services provided by U.S. governments and corporations now assume ubiquitous broadband accessibility, and with this assumption come expectations that cannot be met on tribal lands with the current state of broadband accessibility. As funding and infrastructure licensing processes move online, Tribes are increasingly alienated from the means by which necessary infrastructures are established. In Section 3.1, we show that even with sufficient bandwidth capacity, the wireless links used to extend connectivity over long distances are prone to average packet loss rates of 5%, which degrade media performance in both the

up-link and down-link directions. Packet loss is caused by limitations in 802.11, which make it sub-optimal for transmitting over long-distance wireless links. Even though the WiFi spectrum is inappropriate for the distance requirements of the TDV network, the spectrum is unlicensed and helps reduce the cost of using the network. If Tribes were able to project their sovereignty over radio frequency in tribal land, software defined networks leveraging unlicensed or opportunistically available spectrum could significantly increase the penetration and affordability of broadband services. These spectrum rights have been recommended by the FCC and are currently under discussion as becoming part of spectrum access policy in the U.S. [67].

While it is arguably true that radio spectrum is not optimally utilized or distributed [159], it is a finite resource. In Chapter 2 we demonstrated how quickly infrastructure can become outdated as the size of content continues to grow. Thus, it is imperative to also consider techniques that allow communities to more effectively utilize available infrastructure in such a way that community information goals are maximized.

3.4.3 Conclusion

In this chapter we examined Web usage in two networks operating in rural communities in Indian Country. First, we observed how Web usage in the TDV network was distinct from the U.S. context in which it resides, especially with respect to the prevalence of niche social media traffic, as well as the influence utility infrastructures have on traffic patterns, exemplified in the lack of online shopping traffic due to unreliable delivery infrastructure. Similarly, we identified a high volume of streaming media traffic in the Red Spectrum network as well as a lack of e-commerce traffic. Our work empirically supports trends identified in survey studies regarding broadband usage in Indian Country: we show that social media is one of the most salient applications in the network in addition

to other media-oriented applications like Netflix, YouTube, and iTunes. While packet loss negatively impacts media performance in the TDV network, we use Instagram traffic to identify strong social connections between users within the same reservations and we find that locally created media receives significantly more interactions than non-local media. We also examine Web preference similarities in the Red Spectrum network and find that Web usage patterns exhibited by households are very similar to the aggregate usage patterns of the communities to which they belong. By exploiting these patterns and similarities, architectures that combine local storage, user mobility, and offline social encounters can improve and extend current connectivity.

3.5 Acknowledgements

This work was done in collaboration with Elizabeth Belding, Ellen Zegura, Matthew Rantanen, and Valerie Fast Horse. We would like to thank Joseph Perralta and Geoff Herrin of the Southern California Tribal Digital Village Network and Tom Jones and Justin Hall of Red Spectrum Communications for their assistance with the collection of data used in this analysis. This work was funded by NSF Graduate Research Fellowship Program under Grant No. DGE-1144085 and NSF Network Science and Engineering (NetSE) Award CNS-1064821.

Chapter 4

Social Media Usage in the Tribal Digital Village and Red Spectrum Networks

In this chapter, we analyze social media practices in the TDV and Red Spectrum network to characterize how social media is accessed in these communities. We then distinguish localized usage patterns in the TDV network using data from the `TDV Web-2014` data set. In order to approach social media usage in tribal communities, we pose the following research questions:

RQ1 *Which social media platforms do users in tribal communities use most often?*

RQ2 *What are some of the characteristics of social media usage, including temporal and geographical patterns of usage?*

RQ3 *How does social media usage in a tribal community compare to usage in a non-tribal community?*

Table 4.1: 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

4.1 Overview of Social Media Usage

Our analysis of social media usage in tribal communities begins with an analysis of the social media applications used in both the TDV and the Red Spectrum networks. For this study, we perform a side-by-side comparison of social media usage in both of these networks based on data from the `TDV Web-2017` and `Red Spectrum Web-2017` data sets.

4.1.1 Social Media Applications in the Red Spectrum Network

When examining the Web traffic profile of the Red Spectrum network in Figures **??** and 3.8, 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. We perform a deeper analysis on social media platform usage in the Red Spectrum network by comparing the usage of the top five most accessed social media platforms, reported in Table 4.1.

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 4.1. We find that the average household downloads

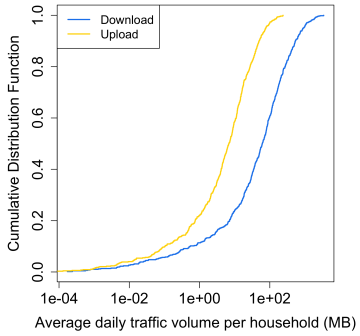


Figure 4.1: Cumulative distribution of the average daily traffic volume for all social media in the Red Spectrum network.

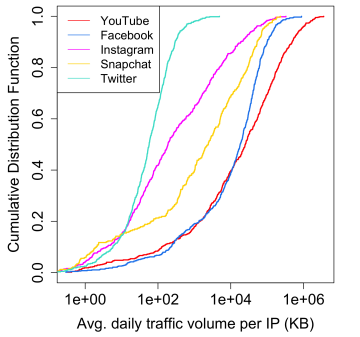


Figure 4.2: Cumulative distribution of the traffic volume generated by the top five most requested social media platforms accessed by users in the Red Spectrum network.

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 greatest traffic volume is associated with YouTube and Facebook followed by Snapchat, Instagram, and Twitter (see Figure 4.2).

4.2 A Study of Instagram Usage in the TDV Network

By the end of 2014, Instagram was among the five most popular OSN platforms in the U.S., with 21% of the entire adult population owning an Instagram account [59]. After Facebook, Instagram has the highest number of users between ages 18-29 and has the highest level of daily engagement (49% of users report interacting with their Instagram account on a daily basis). Moreover, having an Instagram account made users more likely to have accounts on other OSN platform than any other type of account. Considering the general popularity of Instagram in the U.S., the frequency of engagement with the Instagram platform, and our observations of Instagram's popularity in the TDV network, we use Instagram as the lens social media usage in a tribal context.

4.2.1 Initial Examination: `TDV Web-2014`

We begin our analysis with data from the `TDV Web-2014` data set. Table 4.2 provides an overview of the Instagram data we observe in this data set, including information about media interactions and users involved in the TDV Instagram network. In Table 4.2, "content creators" refer to Instagram users who have uploaded media to Instagram. We note that media interactions differ from social interactions: a media interaction includes user-explicit actions such as liking, commenting on, or viewing a media object, while a social interaction only includes user-explicit actions that are announced on the acting user's social feed, such as liking and commenting on a media object.

We find that media views comprise nearly half of all media interactions. This means that only about half of the media interactions that occur in the network are publicly broadcast through the Instagram social network. We also find that only 7% of media

Table 4.2: Overview of TDV Instagram data.

	Total	Local
Media objects	150,368	4,807
Content creators	1,180	164
Media interactions	277,309	19,099
Social interactions	144,721	11,159
Instagram users	NA	238

Table 4.3: Definitions of Instagram interaction types.

Media interactions	User-explicit actions (i.e., liking, commenting on, and viewing a media object).
Social interactions	User-explicit actions that are broadcast to a user's followers (i.e. liking and commenting on a media object).

interactions occur between a user from the TDV network and media created by a user from the TDV network; only 8% of social interactions are between a TDV user and media created by a TDV user. This is consistent with the fact that only 3% of the media objects we observed were created by local users.

At first glance, this low proportion of local interaction seems to indicate a low level of local interest. However, when we look to interactivity based on content creator locale, we find a much higher proportion of social interactions associated with content creators from within the TDV network. We are able to identify the media associated with a content creator based on the media's identifier, which contains a unique ten digit identifier concatenated to the user identifier of its creator. In Figure 4.3, we show the "popularity" of each content creator we observe in the TDV network. Here, we define "popularity" as the total number of media interactions (see Table 4.3) associated with media created by each observed content creator. On average, the TDV community interacts with locally

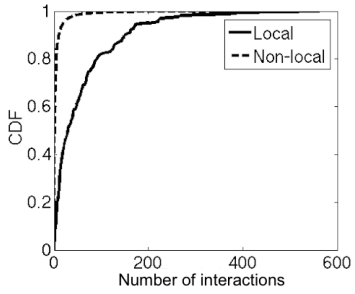


Figure 4.3: Distributions of media interactions associated with local and non-local content creators.

created media 57 times while it interacts with non-locally created media only five times. So while there are far more non-local media objects to interact with, TDV users are much more engaged with content created by local content creators than with content created by non-local content creators.

Based on previous work exploring the connection between social media and cultural resilience in indigenous communities, we study the underlying social structure of the TDV Instagram network [129].We begin by defining a social connection between a given pair of Instagram users u_i and u_j as:

$$C_{u_i} = \sum_{j=0}^{n} P_{u_i}(u_j) \tag{4.1}$$

where $P_{u_i}(u_j) = 1$ if there exists any social interaction between user u_i and media created by user u_j and $P_{u_i}(u_j) = 0$ if no such social interaction exists between user u_i and media created by user u_j.

In Figure 4.4a, we show the number of social connections (C) between each user and all other users in the same reservation (n =$\{11, 17, 25, 69, 114\}$[1] [2]), in the TDV network ($n = 238$), and outside the TDV network ($n = 33,183$). As users interact with other users from broader circles of the Instagram network, the number of social

[1]The value of n is dependent on which reservation u_i is associated.

[2]We exclude users from Manzanita in these calculations as $n = 2$ for this reservation and skews the distribution.

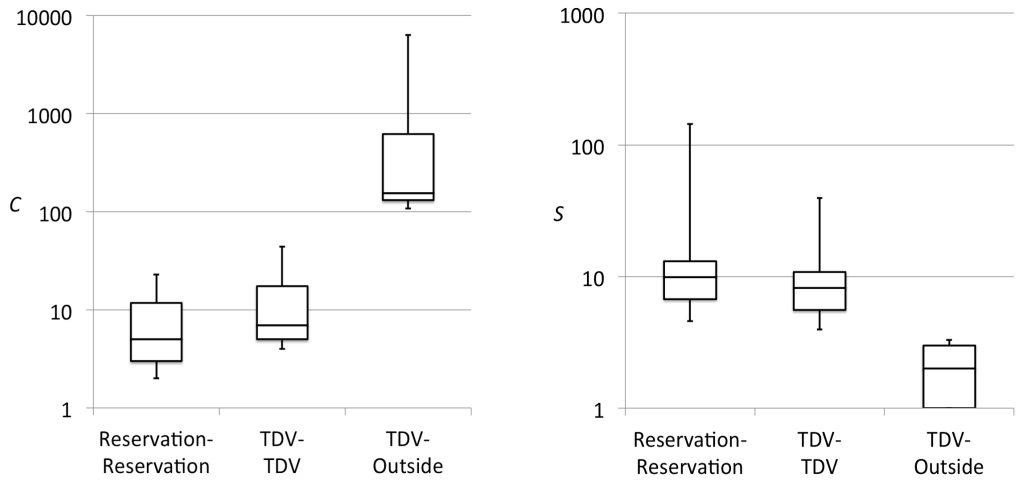


Figure 4.4: (a) Number of social connections per Instagram user and (b) strength of social connections per Instagram user.

connections associated with individual users increases. In addition to the number of social connections, we measure the strength (S) associated with a connection, or the number of social interactions that exist between user u_i and media created by user u_j:

$$S_{u_i,u_j} = I_{u_i,u_j} + I_{u_j,u_i} \tag{4.2}$$

where I_{u_i,u_j} is the number of social interactions between u_i and media created by u_j. Figure 4.4b illustrates the distribution of S for each user in association with other users from the same reservation, from the TDV network, and from outside the TDV network. While we observe a greater number of social connections between local and non-local users in Figure 4.4a, Figure 4.4b reveals that the strength of these connections is weak. In contrast, Figure 4.4b shows that the more proximate users are in the network, the stronger the social connections are between them. Thus, we find that the TDV Instagram network is composed of a dense core of a few strong local connections and expands out via numerous weak connections. Based on this finding, we expect to see a high locality

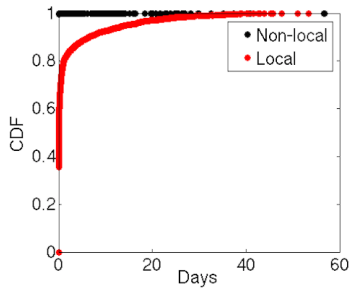


Figure 4.5: Circulation times of local and non-local media.

of interest with respect to media.

In addition to the popularity of media, we examine interactions over time to identify how long media circulates in the TDV Instagram network. We measure this by identifying the time delta between a media object's initial appearance in the TDV network and its final appearance in the network across all media interactions (defined in Table 4.3). We have already established that the number of non-local media far exceeds the number of local media present in the network (see Table 4.2); however in Figure 4.5, we see that locally created content circulates for much longer within the TDV network than non-local media. Looking more closely, we find that 99.6% of the 145,561 non-local media occur only once in the TDV network. Half of the 4,807 local media objects circulate for over 4.2 hours and are liked or commented on an average of 5 times, while 382 local media objects circulate for over a week and are liked or commented on an average of 7 times. Intuitively, longer circulation times should correspond to an increased level of social engagement; however, circulation time is based on media interactions, not just social interactions. This means that media views, which account for 50% of all media interactions, contribute significantly to a media object's circulation time. Therefore, not only is local media more prevalent than non-local media (see Figures 4.3 and 4.4), but it is salient over a longer period of time.

4.2.2 Extending the Analysis: `TDV Instagram-2014`

We extend our analysis of Instagram usage in the TDV network by supplementing data collected using the methods described in Chapter 3 with data culled using the Instagram API. Between June 23 and December 18, 2014, we identified 254 public Instagram user identifiers (unique number corresponding to an Instagram account) by mining the URL field in the HTTP headers captured by the Bro Web logs. For the remainder of this paper, we refer to these public users as TDV users. Because 99% of network traffic is produced at residential access points, we assume that all Instagram users we observe in the network are residents of the 13 reservations serviced by the TDV network. Using the Instagram API, we are able to use public Instagram user identifiers of users in the TDV network to identify the OSN objects they had subscribed to during our collection period. We collect this data in two steps. First, we identify the public user identifiers of public users followed by the TDV users. Users who are followed are called *content creators* for the remainder of this paper. Second, we identify all of the meta data associated with the public media posted by the content creators during our period of interest, including: the media identifier and the creation time of the OSN content. By combining the data collected in these two steps, we generate a final data set that includes: the user identifier of a TDV user, the user identifier of the content creator, the media identifier of the OSN content created by the content creator, and the time the OSN content was created.

In addition, we use "instagram.com/api/v1/media" tags in the HTTP URLs collected in the HTTP Bro logs to identify TDV Instagram users' interactions with Instagram content. These logs include: the timestamp associated with interaction, the media identifier of the OSN content that was interacted with, and the action ("like" or "comment") taken on the Instagram content. By coupling this activity data set with the data set representing the public Instagram content available to each user, we are able to create a combined

Table 4.4: Overview of content interactions observed between June 23 and December 18, 2014.

	Total	Local	% Local
Content objects	1,209,270	3,777	0.31
Content creators	47,645	184	0.39
Social interactions	12,615	1,607	12.7
Instagram users	NA	254	NA

data set that includes: the timestamp of when Instagram content was published, the user identifier of the Instagram content creator, the identifier of the Instagram content object, the proportion of co-located users who have subscribed to the OSN content, and the number of times co-located users interact with the Instagram content.

An overview of the data set we analyze is presented in Table 4.4. We report on the total data set that includes both content generated within the TDV network ("local") and content generated outside the TDV network. Values that are labeled "Local" refer to content objects that were created by content creators from the TDV network. "Content objects" refer to the number of OSN content objects that were published by the content creators followed by public Instagram users in the TDV network. "Social interactions" refer to the number of TDV users' likes and comments on followed Instagram content. Both images and short videos can be published on the Instagram platform; overall, 93.6% of the followed content objects were images and 6.4% were videos.

Locality. Table 4.4 provides an overview of the OSN content objects that were visible to TDV Instagram users during our observation period. We find that less than 0.31% of content objects were created by TDV Instagram (local) users. However, on average, TDV Instagram users interact with locally generated OSN content objects 46.6× more often than non-locally created OSN objects, demonstrating a strong preference for locally created content.

Table 4.5: Definition of terms used to describe properties of OSN content.

Term	Definition
locality	Where in the network with respect to the Internet gateway a piece of content originates. Local content originates within the local subnetwork behind the Internet gateway and non-local content originates from beyond the Internet gateway.
publication time	Time an OSN content object was published to an OSN platform.
coverage	Portion of users who have access to a particular piece of content.
stimulated	OSN content object that has received likes or comments from TDV Instagram users.
dormant	OSN content object that has not received likes or comments from TDV Instagram users.
follow network	All the content creators that a particular user follows, meaning all the content creators from whom a particular user receives OSN content.

Next, we examine locality with respect to users' *follow networks*, or the content creators each user follows. To analyze the impact of locality at finer granularity, we examine interactions between Instagram users connecting from the same reservation. Table 4.6 provides an overview of the number of public Instagram users we observe at each reservation and the corresponding census population. By mapping Instagram users to their corresponding reservation, we are able to calculate the portion of each user's follow network that is comprised of content creators from the same reservation, a different reservation in the TDV network, or outside the TDV network. Given that the majority of observed content creators are from outside the TDV network, it is unsurprising that 97% of users have follow networks that contain more content creators from outside the TDV network

[3]Population sizes from the 2010 U.S. Census. `http://www.sandiego.edu/nativeamerican/reservations.php`

Table 4.6: Population[3] and Instagram statistics for six TDV reservations.

Reservation	Instagram users	Population size
Pala	71	1,573
Rincon	61	1,495
San Pasqual	54	752
Mesa Grande	48	75
Pauma	17	186
Manzanita	3	69

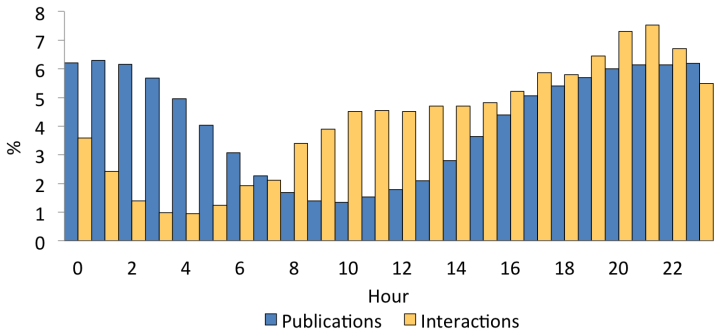
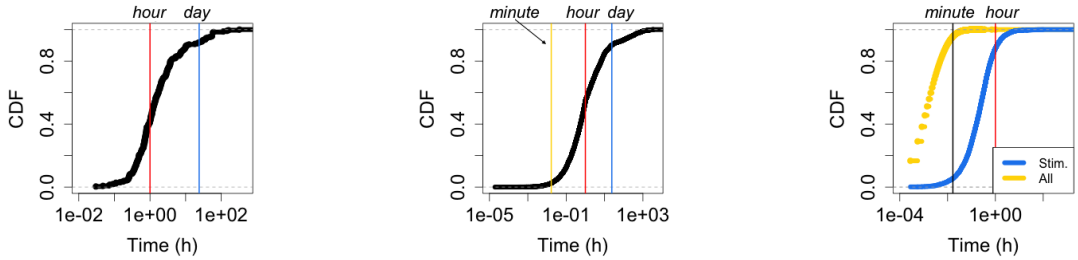


Figure 4.6: Percentage of overall activity that occurs per hour.

than from inside the TDV network. When considering only TDV users that comprise a given user's follow network, the median percentage of content creators from the same reservation is 2.9× the median percentage of content creators from another reservation in the TDV network. This indicates the potential impact that community boundaries have on the success of sharing limited bandwidth. The broader the boundaries that define a "co-located" group of users, the less likely they are able to capture meaningful connections between that set of users.

Publication Time. By understanding how frequently OSN content is published, received, and interacted with, we develop a temporal sense for how frequently OSN content broadcasts would need to occur in order to sustain a relevant information flow across a shared low-bandwidth link. While we observe significantly more publications than interactions (see Table 4.4), we normalize the frequency of publication and interaction



(a) (b) (c)

Figure 4.7: Distribution of the time interval between (a) received OSN content objects, (b) publication time and time of first interaction, and (c) publication times between incoming OSN content objects.

activities by reporting the percentage of activity that occurs per hour in Figure 4.6. We note that the percentage of activity is based on all activities observed during our six month period of interest. As shown in Figure 4.6, the majority of publications occur during the evening and early morning, while the frequency of interactions is consistent with daily Web traffic patterns observed in Chapter 3.2.1.

To understand post frequency from the perspective of the user, we examine the distribution of the average number of hours between each user's reception of a new OSN post in Figure 4.7a. This relates to how frequently a user's newsfeed is updated with new content. The median frequency of reception experienced by users is one received post per 1.23 hours; 40% of users receive a new post more frequently than once per hour. We observe the number of hours between the initial publication of a post and its first TDV-originated interaction (like or comment made on the post by a TDV Instagram user) in Figure 4.7b. The mean number of hours between a post's publication and the initial interaction on the TDV network is 1.04 hours; 90.2% of posts that are interacted with receive their initial interaction within 24 hours of publication. We take these values into consideration when evaluating potential approaches for scheduling OSN content broadcasts over RBDS.

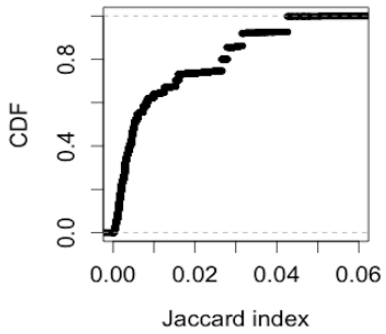


Figure 4.8: Distribution of the Jaccard similarity indices associated with pairwise comparisons of each user's follow network.

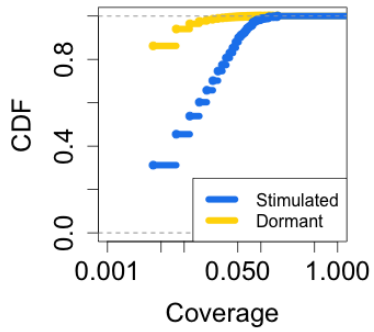


Figure 4.9: Distribution of coverage provided by stimulated and dormant content.

In order to conceptualize how the rate of OSN content publication would impact the bandwidth required to update TDV OSN users in real-time, we examine the distribution of time intervals between all viewable posts assuming a naive First Come, First Serve (FCFS) approach. In Figure 4.7c, we plot the time interval between all unique viewable content in yellow (note the logarithmic x-axis). We find that the rate of content publication translates into a median interval of 2 seconds between OSN content posts that would traverse a shared link. Depending on the size of the post, the observed data rate could require bandwidth capacities of 2.24 kbps for the most bandwidth-light OSN platforms

(Twitter) or beyond 80 kbps for media-based OSN platforms (Instagram), assuming an average data size of 560 bytes and 20 KB, respectively. Since the RBDS protocol operates at only 1.1875 kbps, transferring all OSN content on the current FCFS basis is clearly infeasible. However, there are reasons to remain optimistic about transferring OSN content to disconnected communities via RBDS. Only 0.55% of all viewable published content receives interactions from TDV users. We plot the distribution of time intervals between the publication times of stimulated OSN content posts using FCFS scheduling in Figure 4.7c as a blue line. For the stimulated OSN content objects, the median interval between publication times of OSN content is 13.66 minutes. This rate of publication translates into bandwidth requirements of 0.00546 kbps for the most bandwidth-light platforms (Twitter) and approximately 0.195 kbps for media-based OSN platforms (Instagram). The lower data rates suggest the feasibility of transmitting the most relevant media over RBDS. The challenge lies in the identification of relevant media, which we discuss in the following section.

Coverage. We next investigate whether the high locality of interest identified in our previous study of the TDV network will be reflected in the amount of overlap observed between TDV Instagram users' follow networks. Using the Jaccard index [102],

$$jaccard(A, B) = \frac{|A \bigcap B|}{|A \bigcup B|} \tag{4.3}$$

where A is the set of users that comprise one user's follow network and B is the set of users that comprise another user's follow network, we compare users' follow networks and report the distribution of similarity indices in Figure 4.8. Contrary to our expectations, the Jaccard similarity between users' follow networks is quite low, with similarity indices ranging from 0 to 0.051 with a median similarity index of 0.0052.

Patterns of interaction with OSN content provide nuance to our understanding of

coverage and relevance. As illustrated by Figure 4.9, we observe a significant difference ($p < 2.2 \times 10^{-16}$) between the coverage associated with stimulated content and dormant content. While stimulated content does not represent the entire set of content that would be considered relevant to the TDV users, the higher coverage values associated with stimulated content signifies a potential heuristic for determining content relevance to the group of TDV users and can be used as a prioritization value for scheduling content broadcasts over RBDS. We explore this in more detail in the following section.

4.3 Discussion and Conclusion

Work in the social sciences has identified the importance of online social networks (OSNs) to Tribal identity and cultural preservation [129]. In particular the formation of "bonds" and "bridges" within Tribal OSNs has significant impact on the strength and resilience of indigenous culture [135]. Bonding relations within a community help members make sense of negative experiences and look forward to a more positive future, creating cultural continuity that connects the traditional culture of the past to the development of a new culture of the present. Bridging connections between different communities allows for communal empowerment and influence on the wider society. In Section 4.2, we investigated various dimensions of local versus non-local media prevalence and propagation in the TDV Instagram network. With respect to media interactions on Instagram, our findings show that there are many weak bridging connections and a dense core of strong bonding connections. This leads us to believe that social media in Tribal functions to strengthen and maintain bonding connections; however, it is unclear whether bridging connections exert any quantifiable influence over outside communities.

The high locality of interest with respect to locally created media presents opportunities to improve bonding connections in several ways. Current connectivity can be

improved by storing locally created social media content locally, which would prevent users from uploading media over lossy relay and access links that extend upload completion times, improving upload success rates. This strategy would also improve download performance (particularly for videos), as user quality of experience could be improved by avoiding lossy links. Moreover, if locality of social media interest corresponds to spatial closeness, social bonds could be used as a means of extending connectivity via opportunistic encounters with peers. This way, users who have access to spaces of ubiquitous connectivity might serve as media vectors, transferring relevant media over Bluetooth, WiFi Direct, or NFC to other users as they come into contact geographically.

4.3.1 Conclusion

Prior studies investigating the informational needs of reservation communities have emphasized the importance of OSN content for connecting users together through indigenous language [129, 76], cultural practice [31], experience sharing [31], participation in political and social movements [6, 125], and daily communication [129]. Understanding the information habits of indigenous communities enables researchers to develop new solutions that help users that live beyond the borders of Internet service connect to each other through existing platforms for computer-supported collaboration (e.g., OSNs). Our characterization of OSN usage by TDV Instagram users reveals a high locality of interest that is consistent with reports of OSN usage in both indigenous [129, 31, 76] and rural [106, 77] communities around the world. Our analysis of social media usage in the Red Spectrum network reveals that Facebook (facebook.com and fbcdn.com) represents the most requested domain in the network and Twitter is the most prevalent social media platform with respect to percentage of subscribers who access. An examination of the top Web domains in the TDV network during the same observation window reveals that

Instagram remains the most accessed social media platform in that network. In this chapter, we demonstrate the network mechanisms by which information is distributed using methodologies that are agnostic to the content itself. In the next chapter, we examine how some of these social media structures form and interact with respect to information as it is characterized by its associated media and content.

4.4 Acknowledgements

This analysis was done in collaboration with Elizabeth Belding, Ellen Zegura, and Matthew Rantanen. This work was funded by NSF Graduate Research Fellowship Program under Grant No. DGE-1144085, the NSF Network Science and Engineering (NetSE) Award CNS- 1064821, and the NSF Networking Technology and Systems (NeTS) Award NSF-1563436.

Chapter 5

Interdisciplinary Approach to Understanding Native American Political Discourse on Social Media

In the previous two chapters, we analyze how users in two different tribal communities use Internet connectivity with an emphasis on Web usage and social media usage. Motivated by calls issued by the U.S. Government Accountability Office to connect Internet connectivity and political action in Indian Country, we examine how social network structures associated with Native American advocates form and interact with political action content.

As highlighted in Chapter 2.1.1, Native Americans comprise an exceptional class of citizenship within the U.S. While many Native Americans are voting members of tribal nations, they are also eligible to vote in local, state, and national elections. However, the historically agonistic relationship between the U.S. federal government and Native American nations, has discouraged Native American individuals from engaging with electoral politics in the U.S. [87, 170, 212]. Moreover, Indian Country, which is associated with

some of the largest Native American voting blocs, suffers from a lack of communications infrastructure, limiting Native American individuals' potential for political engagement through digital means. To demonstrate the role that Internet infrastructure plays in the empowerment of marginalized communities, we examine how the Twitter platform is used to disseminate political action content and as a space for Native American advocates and their followers to congregate around this content.

To the best of our knowledge, there have been no network scientific studies of Native American political engagement through social media, although social media uses are observable in Native American policy arenas [14, 55, 105, 124, 207, 109, 126, 75]. Indeed, presidential hopeful Bernie Sanders hired two well-known Native American rights advocates to help craft his social media campaign [120, 131]. The 2016 U.S. presidential elections present a unique opportunity to investigate the political content propagated by Native American advocates[1] representing a diverse swath of Indigenous and tribal interests. As a cursory investigation into the characteristics of Native American political content on Twitter, this research asks:

RQ1 *What political content do Native American advocates share on Twitter?*

RQ2 *What are the network characteristics of sub-communities present within the Twitter streams of Native American advocates?*

RQ3 *In light of bandwidth restrictions in Indian Country, what are the bandwidth characteristics of content propagated by and from Native American advocates?*

To accomplish this, we worked with Indigenous scholars and community-based activists to curate a list of the Twitter hashtags and user accounts they follow to share

[1]We use the phrase *Native American advocate* to refer to activists, journalists, newsgroups, scholars, and non-governmental organizations that represent North American Indigenous peoples, nations, and individuals.

political information. We culled the most frequent hashtags and top user accounts to generate a data set. We collected and characterized 11,102 tweets generated and/or shared by Native American advocates active on Twitter. We contrast our findings as they pertain to the Twitter activities of Native American advocates to 46.5 million tweets sampled from the general Twitter feed. Using the social connectivity information embedded in the Native American advocates data set, we identify network sub-communities, and highlight ways that dispersed efforts pull from similar bases of support, ultimately providing a characterization of Native American and Indigenous political agendas as manifested by advocates online. Finally, we discuss these findings in the context of on-the-ground realities for Native American people.

5.1 Theoretical Framework

Our work is best understood through the combination of two theories: connective action and media richness.

5.1.1 Connective Action

For heterogeneous Native American groups, hosting political "diffuse conversations [19]" through social media contributes to media salience in spite of mass media marginalization of Native American political issues [188]. Agarwal et al. examine the role of Twitter in the Occupy movement, using the constant comparative method and with empirical analysis of network artifacts (specifically, tweet records collected using the Twitter Streaming API) [1]. Characterizing the networks that allow for successful Internet-based SMOs, Gloor defined Collaborative Innovation Networks (COINs): "a cyberteam of self-motivated people with a collective vision, enabled by the Web to collaborate in achieving a common goal by sharing ideas, information, and work" [78]. Similarly, through analysis

of the online uses of collective action networks–brick-and-mortar institutions and face-to-face groups–Bennett and Segerberg identified *connective action networks*, groups of individuals who may only encounter each other through online spaces and who are unaffiliated with SMOs or brick-and-mortar institutions, yet who mobilize toward common goals [15]. In that sense, we approach Native American advocates' uses of social media as the connective tissue binding multiple political action environments, where actors who may or may not know each other and who may or may not belong to SMOs nevertheless agree to propagate content and adopt discourses related to certain issues.

5.1.2 Media Richness

Daft and Lengel's Media richness theory (MRT) defines *media richness* as "the relative ability of information to influence or change mental representations and thereby to facilitate learning [45]." More specifically, the richness of each medium is based on the following: "the use of feedback so that errors can be corrected; the tailoring of messages to personal circumstances; the ability to convey multiple information cues simultaneously; and language variety [116]." As an online social media platform, Twitter is capable of posting users' content and responses to content. The platform supports embedded multimedia content including audio, photo, and video in addition to 140 characters of text. While Twitter operates as a broadcast medium where tweets are visible to any other user on the platform, tweets can be personalized and directed leveraging usertags and hashtags. In particular, we interpret the impact of media richness from the perspective of Twitter audiences connecting from Indian Country, where persistent digital divide effects likely impact the Internet access and connectivity capacities of some of the largest Native American voting blocs.

5.2 Methodology

By applying decolonizing and post-structural methods with network analysis, we follow an approach similar to Garrido, who, in tracing the Zapatista movement, reconstructed a network from digital artifacts, applied categorizations based on domain knowledge, and used network analysis to determine relationships among actors and topics in the network [75]. Thus we surface, describe, and quantify what Smith refers to as the "[networking] process which Indigenous peoples have used effectively to build relationships and disseminate knowledge and information" [178].

5.2.1 Statement of Positionality

The research team consists in part of Native American advocates and educators with a combined record of over 20 years of experience working with Native American SMOs, cultural revitalization efforts, and Native American political theorists. Results that pertain to political engagement are analyzed from within an Indigenous political science paradigm, which is premised on the assumption of colonizing logics in modern Westphalian nation-states, the politics of recognition, and theories of tribal sovereignty and self-determination [44, 61, 155, 188].

For Native Americans, sovereign rights refer to the rights they bear within their tribe as it is recognized by the U.S. federal government. At present, there are 586 federally-recognized tribes within U.S. borders, and U.S. congressional representatives acknowledge tribal rights occasionally, and not as a matter of course. Many social scientists interpret Indigenous peoples' social movements as entirely identity-based movements, which is technically erroneous, as many Indigenous peoples' movements are also expressions of the sovereign autonomous rights of federally-recognized tribal governments [188]. This study is designed to reveal the propagation of Native American political content through

network analysis of Twitter data sets, with conscientious regard to issues affecting Native American peoples, given their limited access to the Internet, limited resources for Internet infrastructure innovation, and the constraints of democratic political participation for Indigenous peoples.

5.2.2 Definitions

For this study, we defined *political content* as data exchanged over technical network channels that pertains to political engagement. We define *political engagement* as human activities that contribute to awareness of justice in governmental affairs and moral or ethical behavior of government officials or institutional authorities, mostly as consciousness-raising, protest strategies, or sustained critique. Political engagement can include *political action*, which refers to direct and indirect methods that individuals utilize to change a governmental status, including political participation through voting, registering to vote, donating to campaigns, and petitioning.

We define *content propagation* as the transmission of data across the Twitter platform through the intentional user techniques of tweeting, retweeting, and embedding artifacts (i.e. photos, videos, and URLs) into the tweet. We also distinguish the *actor* from the *user*, in which the actor is a human or non-human node in a time interval within a network map, while a user is an individual or organization with a unique identifying Twitter account. Additionally, we define *sub-communities* as clusters of actors identified using the Louvain measure of density of Jaccardian similarity edges between hashtag-centric ego networks.

5.2.3 Data Curation

Functioning as participant observers, the research team created a list of search terms based on their own Facebook newsfeeds, groups, and friends lists and Twitter streams. In addition, the team reached out to fifteen associates, also Native American activists, advocates, educators, and journalists, who likewise contributed Twitter hashtags and usernames. The team queried individuals across a range of advocacy roles, from individuals in prominent institutional policy roles to individuals working in remote reservation areas and focusing on local issues. The process of manual curation resulted in a list of 45 hashtags and 33 user accounts.

5.2.4 Data Collection

Between February 11 and March 31, 2016 (during the height of the U.S. presidential primary election season), the team queried the Twitter Streaming API using a list of 45 hashtags and 33 user accounts, specifically tracking the number of original tweets, retweets, and users associated with these hashtags and user accounts. We provide an example to demonstrate how our query methodology functioned. Using the hashtag *#mmiw*[2], we captured all original tweets and retweets containing the string "mmiw." The Twitter Streaming API is not case sensitive, so all possible letter-case combinations (e.g. "MMIW," "Mmiw," "mmiW") were captured in our sample.

One of the limitations of the Twitter Streaming API is that it does not allow API users to filter by specific hashtags, meaning the hashtag symbol ("#") is ignored in the query. Thus, the API interprets the hashtag as a keyword. This becomes problematic when filtering for acronym hashtags such as *#mmiw* because they can be matched to tweets in non-English languages. In order to ensure that the tweets used in our analysis

[2] "Missing and murdered Indigenous women."

	Native American Advocates	General
Total tweets	11,102	46,495,733
Unique tweets	5,172	24,619,723
Retweets	5,930	21,876,010
Users	5,019	13,879,253
Content creators	2,086	3,064,395

Table 5.1: Overview of the Twitter data sets collected between February 11, 2016 and March 31, 2016.

reflect our targeted hashtags, we imposed our own post-filtering process. This process includes translating the original tweet text to a lowercase string, parsing the string into whitespace-separated tokens, then using regular expression matching across each token to assess whether any of the desired hashtags were included in the text of the resulting tweet or retweet. All tweets that included at least a single match with a hashtag in our list of hashtags are included in the data set.

The details of this data set are presented in the "Native American Advocates" column of Table 5.1. In order to provide a larger context for the Native American advocates data set, we use the sampling mechanism of the Twitter Streaming API to procure a data set that represents a random 1% sample of all tweets generated between our study dates of February 11 and March 31. We filter this data set down to English-language tweets and report the details of the data set in the "General" column of Table 5.1. The original JSON files collected for this research, query terms used to seed the Twitter Streaming API, and software used to collect and analyze data are available for public access at `https://github.com/mvigil90/IndianCountryTweets`.

5.2.5 Data Analysis

The Native American advocates data set resulted in a list of 5,019 users, including the preliminary 33 recommended user accounts. The 5,172 unique tweets (not includ-

ing retweets) were generated by 2,086 users, or content creators. The data set consists of 11,102 total tweets. We apply our combined experience with Native American political issues to identify the topics in the most frequently propagated content. Cursory qualitative review of randomly selected subsets of this data set includes topics such as: news about missing Indigenous women, presidential campaign messaging, notices about environmentally damaging projects, and updates about the Indian Child Welfare Act.

We also examine the types and sizes of media embedded within the tweets we collect. We identify tweets with embedded media as those that contain the full URL associated with linked media. We label each tweet record as containing either an embedded photo, embedded video, or no media. We discern photos by searching the embedded media URL for the substrings associated with embedded image types on Twitter, namely PNG (".png") and JPEG (".jpeg"). We also discern tweets that link to videos by searching the tweet body for regular expressions mentioning videos ("video") or containing URLs to popular video sites ("youtube.com", "vimeo.com", and "vine.co"). Finally, we examine the sizes of the embedded content. To do this for photos, we use cURL[3] to download the photo from the URL embedded in tweets and ascertain the file size. For videos, which are streaming content, we first manually identify the temporal length of each video and then find the file size by multiplying the video time by various data rates that Twitter supports.

Finally, we applied network analytic approaches–specifically social graph analysis, descriptive statistics, sequence analysis, and cluster analysis–to characterize network structures in the Native American advocates data set. We do this according to the process outlined in Figure 5.1: (i) We create hashtag-centric ego networks (H_i) where hashtags (h_i) represent ego nodes and the actors (a_i) that tweet and retweet a hashtag are the neighbor nodes, then (ii) we use the Jaccard index to calculate how similar each hashtag-

[3]cURL. `https://curl.haxx.se/`

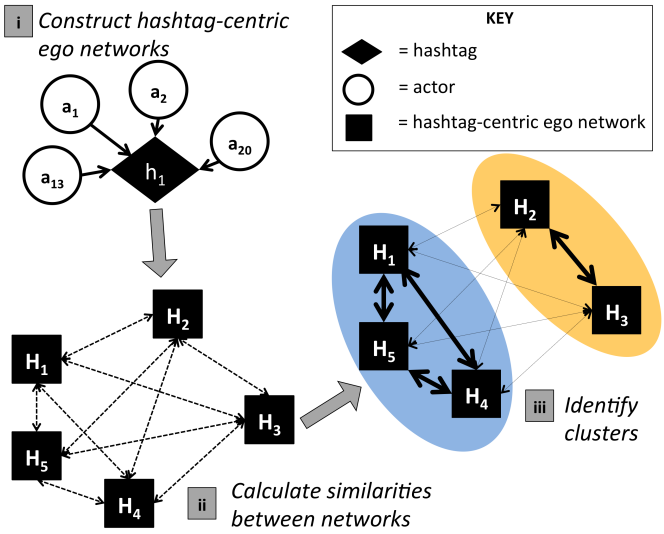


Figure 5.1: Network analysis methodologies used to identify sub-communities in the Native American advocates data set.

centric ego network is to each other [102], and finally (iii) we use the Louvain method to identify sub-communities of actors that tend to form around clusters of hashtags [17].

5.2.6 Critique of Methodology

In a recent study, Tufekci asserted a number of methodological and inference issues commonly associated with social media big data analysis including: limited platform representation, selection on dependent variables, unrepresentative sampling, ignorance of wider social ecology of interaction, ambiguous interaction sentiment, disparity between actual and theoretical usage, inappropriate application of network methods, ignorance of field effects, and skews caused by human self-awareness [186]. Similarly, Morstatter et al. presented a critique of the sample quality provided by Twitter's Streaming API [130]. Here we provide a critique of our collection methodology in light of the most common methodological criticisms.

Representation. Although the data sets resulting from our collection methodologies are not representative of Native American social media activity as a whole, it does

represent the social media interactions between a collection of Native American advocates, Native American political issues, and users who follow them. As a highly interdisciplinary team representing research expertise in computer science, Indigenous information systems, and Native American public policy, we have thoughtfully applied domain knowledge in discerning methodologies for data collection and analysis. We deliberately limited our analysis to a single social media platform for two reasons. Our initial objective was to investigate directed information propagation patterns that occurred between Native American advocates, their audiences, and their information sources. Twitter proved to be the best platform for observing this type of propagation. Second, there was a lack of data from the outset regarding Native American online political engagement and it made sense to begin with a platform that enabled public visibility to a wider portion of user accounts and content [81]. There is indeed a greater social ecology not fully captured by the study of a single platform; however, when attempting a cursory investigation into the relationship between network infrastructure and the propagation of minority perspectives online, it is reasonable to begin by understanding interactions as they take place over a single platform.

Sampling. Morstatter et al. suggest mitigating the sampling effects in Twitter's Streaming API by generating more specific parameter sets with different users, keywords, and geographical bounding boxes [130]. In our own methodology, we attempt this in several ways. First, we curated as specific a list of hashtags and user accounts as possible using our connections to Native American advocates, then queried the Twitter Streaming API using two different application keys, one for hashtags and the other for user accounts (generating two overlapping samples that are $\leq$1% of all simultaneous Twitter activity). Our final Native American advocates data set is a union of these two data sets. While it was possible for us to impose geographical restrictions on the samples, we decided to forgo these restrictions for two reasons. First, the bounding box we required to capture

Tag	Original tweets	Retweets	Users
#Indigenous	2,303	3,042	2,839
#mmiw	607	1,054	1,031
#tairp	358	987	527
#nativelivesmatter	311	278	305
#nativeamerican	205	85	97
#idlenomore	189	199	193
#ndn	184	51	66
#hiring	177	1	9
#colonialism	151	136	259
#cdnpoli	140	176	186

Table 5.2: Top 10 most posted hashtags in the Native American advocates data set.

tweets about Native American political issues was too large to function as a practical filter. Second, because Native American and Indigenous political engagement explicitly revolves around transnational sovereign relationships between nations, studies of Indigenous political engagement are not bound by state borders, but rather are shaped by Indigenous social and cultural practices, issues, and affiliations that occur in the margins of national and state borders [188].

5.3 Native American Political Content on Twitter

Our first research question investigates the types of content Native American advocates post on Twitter and contrasts this content to content present in the general Twitter stream. We also examine the Native American advocates data set for content pertaining to political action and compare it to a general data set.

5.3.1 Hashtags

We begin our analysis of topics by examining the hashtags associated with the posts in our Native American advocates data set. Overall, we observe 2,885 unique hashtags. We report the top 10 most posted hashtags in Table 5.2. While many of the top tags pertain to Indigenous and Native American identity (*#Indigenous*, *#colonialism*, *#tairp*[4], *#nativeamerican*, and *#ndn*[5]), a few of the tags represent specific causes, including femicide awareness (*#mmiw*) and murder/suicide awareness (*#nativelivesmatter*). For the top hashtags that correspond to identity, we find that they are more likely to be paired with other hashtags than to be used as standalone hashtags ($P(c_t > 1|t_{identity}) = 0.90$ where c_t represents the number of hashtags associated with a tweet, t). This is in contrast to the top hashtags that emphasize specific issues, which are more likely to exist as standalone hashtags ($P(c_t > 1|t_{issues}) = 0.42$).

5.3.2 Categories

Next, we examine the top 100 most frequently occurring hashtags in both the Native American advocates and general Twitter data sets and categorize them with one of the topic labels described in Table 5.3, which represent: identity (I), civil rights (CR), current events (CE), environmental issues (EI), and other (O). We note that for the general data set, the identity category refers to hashtags that signify and promote group identity. We then report the percentage of tweets associated with each of the top 100 hashtags that fall into each category in Figure 5.2. For the Native American advocate data set, the category with the most associated hashtags is the identity category (I) with 55% of the hashtags, followed by the civil rights category (CR) with 32% of the hashtags. In contrast, the top category for the general data set is other (O) with 59%, of which 69%

[4] "The American Indian Red Power."
[5] "[American] Indian."

Topic	Description	Examples
I	Relevant to Indigenous, Aboriginal, or Native American peoples and promotes Indigenous identity through acknowledgement of Indigenous language, art, culture, and education.	*#Indigenous*, *#ndn*, *#nativeamerican*, *#metis*
CR	Promotes political and social justice for minorities, particularly rights while engaging with law enforcement and the legal system.	*#nativelivesmatter*, *#mmiw*
CE	Highlights news events or campaigns that occur during or near the observation window.	*#nativevote16*, *#nativesforbernie*, *#caucus*
R	Points to resources including job advertisements and health services.	*#ihs*, *#jobs*, *#hiring*
EI	Related to environmental issues and concerns, either current or longstanding.	*#pipeline*, *#saveoakflat*, *#climatechange*
O	Miscellaneous tags that do not fit into the above categories.	*#love*, *#facebook*

Table 5.3: Description and examples of topical categories that are applied to the top 100 hashtags in each data set.

pertain to popular awards (e.g. Nickelodeon's Kids' Choice Awards or the iHeart Radio Music Awards) and entertainment. The identity category in the general data set has 22% of the top hashtags and 97% of these refer to celebrity fan bases. It is also worth noting that the general data set contains no civil rights or environmental hashtags in the top 100 hashtags.

5.3.3 Circulation

In addition to overall tweet count, we evaluate the circulation of topics in the data sets by examining the prevalence and persistence of hashtags. Defined by Paxson when characterizing the presence of routes in the Internet, prevalence and persistence are metrics that can be used generally to characterize churn [151]. Churn refers to the levels of instability surrounding a hashtag or network structure that manifests in the flow of information. In Figure 5.3, we show the distributions of prevalence over one-day intervals

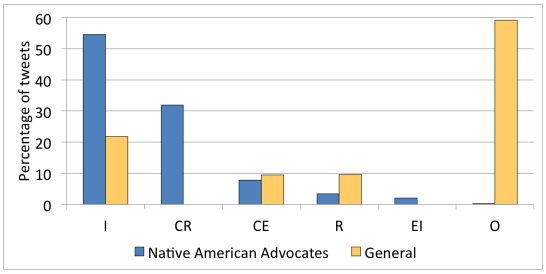


Figure 5.2: The percentage of tweets from the Native American advocates data set that fall into each topical category defined in Table 5.3.

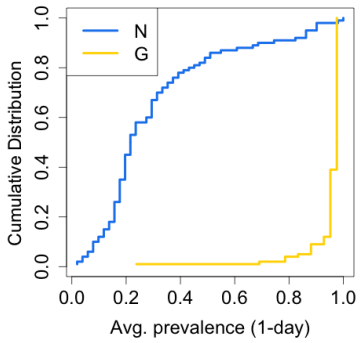


Figure 5.3: Cumulative distribution of hashtag prevalence over day-long intervals where "N" corresponds to the Native American advocates and "G" represents the general data set.

for the top 100 hashtags in the Native American advocates and general data sets. In this context, we define *prevalence* as the portion of day-long time segments in which a hashtag is present relative to all time segments. For the Native American advocates data set, the median prevalence is 0.22, meaning that at least half the hashtags are present across 22% of the days contained in our observation period. Only 5 hashtags are present for more than 90% of the observation period. These include three hashtags that denote Indigenous identity (*#Indigenous, #nativeamerican*, and *#tairp*), one hashtag raising awareness for violence against Native women (*#mmiw*), and *#jobs*. On the other hand, the median prevalence for the top 100 general hashtags is 0.97. Thus, it is apparent that identity-based hashtags are the ones with the greatest longevity in the Native American advocates data set, whereas most hashtags in the general data set are highly prevalent

Time scale	%	Notes
minutes	5.2	"Ephemeral." These hashtags are retweeted for only minutes after the original post and represent transient topics.
hours	0.8	"Event-driven." These topics represent reactions to specific events and headlines.
days	2.9	"Recurrent." These topics represent recurrent issues and causes in the Native American advocates data set.
all	1.7	"Pervasive." These are topics pervasive to the Native American advocates data set.

Table 5.4: Summary of topic persistence at different time scales for Native American advocates data set.

(meaning most hashtags appear in each day of our sample window–see Figure 5.3). An explanation for this is the much larger volume of tweets in the general data set and the fact that a significantly larger portion of the population is represented in that data set.

We next characterize the persistence of each hashtag. We define *persistence* as the number of consecutive time segments in which a hashtag appears. For the Native American advocates data set, we examine persistence at the magnitude of minutes, hours, and days. We calculate persistence by identifying the initial appearance of a tag and counting the consecutive time intervals at which it is present. For hashtags that appear more than once, we report the average persistence across all appearances. In Table 5.4, we provide an overview of the percentage of hashtags that are persistent at each time interval, meaning they are present in the data set for more than one consecutive interval. In addition to the interval categories, we add a category that corresponds to content that is persistent across all time scales. Content that is persistent only at the scale of minutes is

Native American Advocates				General			
Username	Tweets	Users	Prevalence (%)	Username	Tweets	Users	Prevalence (%)
@POTUS	39	67	29.4	*@realDonaldTrump*	75,629	63,237	97.6
@BernieSanders	24	38	35.3	*@tedcruz*	27,420	23,729	97.6
@zhaabowekwe	15	25	21.5	*@HillaryClinton*	21,356	24,704	97.6
@HillaryClinton	10	3	3.9	*@BernieSanders*	20,831	21,507	97.6
@goldmanprize	5	3	1.9	*@marcorubio*	12,885	12,997	95.2
@indiancountry	7	6	7.8	*@FoxNews*	11,172	11,958	95.2
@SenSanders	6	8	9.8	*@POTUS*	9,048	13,174	97.6
@BarackObama	6	11	7.8	*@YouTube*	7,589	7,490	95.2
@WinonaLaduke	5	22	13.7	*@CNN*	6,883	8,518	95.2
@realDonaldTrump	5	6	5.8	*@JohnKasich*	5,877	7,140	95.2

Table 5.5: Statistical overview of the most mentioned users in political action tweets.

classified as "ephemeral" content. The ephemeral hashtags with the longest persistence were associated with Bernie Sanders (*#wearebernie*, *#tulalipforbernie*, and *#bernieinseattle*), larger social justice movements (*#brownlivesmatter* and *#seniorcitizens*), and calls to mobilization (*#urgentaction* and *#sign*). The most persistent ephemeral hashtag (*#wearebernie*) lasted for 6.5 consecutive minutes and was retweeted 31 times. We also identify hashtags that are persistent on the order of hours, or "event-driven" hashtags. These account for a small portion of the hashtags, and are exemplified in posts that refer to specific events, including Internet Friendship Day (February 13), the death of U.S. Supreme Court Justice Antonin Scalia (February 13), and a march to raise awareness for missing and murdered Indigenous women (February 14). The most persistent of the event-driven hashtags (*#mmiw*) lasted 2 hours and was retweeted 7 times. "Recurrent" hashtags correspond to persistence on the order of days. These hashtags are associated with more intersectional Indigenous concerns, most predominantly violence against women, environmental issues, and cultural appropriation. The most persistent of these hashtags (*#mmiw*) lasted for 3 consecutive days and was retweeted 8 times. Finally,

Keywords	bernie, bern, sanders, hillary, clinton, barack, obama, donald, trump, cruz, rubio, kasich, senator, president, caucus, primary, democrat, republican, ballot, vote, debate, register, convention, elect, incumbent, poll, political, politics, gop, liberal, conservative, congress, potus, supreme court, senate, representative, delegate

Table 5.6: Keywords used to identify political action content.

we examine hashtags that are persistent at all intervals (minutes, hours, and days). We refer to these hashtags as "pervasive," since the issues they address represent some of the most ubiquitous topics we encounter. Pervasive hashtags are predominantly associated with Indigenous and Native American identity, femicide, murder/suicide awareness, and political action. The most persistent of the pervasive hashtags (*#Indigenous*) lasted 49 consecutive days. At the scale of hours, it persisted 7 hours and at the scale of minutes it persisted 2.54 minutes. Overall, it was retweeted 3,042 times.

5.3.4 Political Action Hashtags

We identify hashtags in the top 100 that are related to political action. For the Native American advocates data set, 8 of the top 100 hashtags are associated with political action, including: *#cdnpoli*[6], *#auspol*[7], *#fnpoli*[8], *#feelthebern, #wearebernie, #natives-forbernie, #nativevote*, and *#nativevote16*. The median day-long prevalence for these hashtags is 0.33 ($\sigma = 0.25$). For the general data set, 7 of the top 100 hashtags are associated with political action. These include: *#trump2016*, *#trump*, *#feelthebern*, *#cruzcrew*, *#gopdebate*, *#demdebate*, and *#pjnet*[9]. The median prevalence for these

[6] "Canadian politics."
[7] "Australian politics."
[8] "First Nation politics."
[9] "Patriot Journalist Network."

hashtags is 0.98 ($\sigma = 0.01$). As mentioned previously, an explanation for the significant difference in political action hashtag prevalence is the fact that the general data set is much larger than the Native American advocates data set. When comparing the top hashtags in these two data sets, it is also worth noting that both Democratic and Republican presidential candidates and debates are represented in the general data set. This is a contrast to the Native American advocates data set where the top political action hashtags are used in tweets that are non-opinion bearing statements (typically associated with news sources and voter registration campaigns) or highlight Democratic presidential candidate Bernie Sanders.

5.3.5 Political Action Content

One of the distinguishing characteristics of our data set is the fact that it was collected in the midst of the 2016 presidential primary election season in the U.S. Knowing this, we filter tweets that contain keywords associated with political action (see Table 5.6).

We identify 528 unique tweets (938 total tweets) in the Native American advocates data set that contain these keywords, which represents 10.2% of all 5,172 unique tweets we observe in the data set. In comparison, we identify 2,063,583 unique tweets (2,919,275 total tweets) that contain these keywords in the general data set, which represents only 5.7% of all 3,608,8642 unique tweets we observe in the general data set.

We begin our analysis of political action by examining the top 10 most frequently mentioned users in the subset of tweets that match keywords in from Table 5.6. In Table 5.5, we report the number of tweets that mention a username, the number of unique users who tweet or retweet posts mentioning a username, and the prevalence of the mentioned username (on a one-day scale). Of the top 10 most mentioned usernames for the Native American advocates data set we observe 6 individuals, 1 NGO, and 1 Na-

tive American news network. With the exception of two, the individuals represented in Table 5.5 are all politicians. This includes current U.S. president, Barack Obama (*@POTUS*[10] and *@BarackObama*), as well as current U.S. presidential candidates: Bernie Sanders (*@BernieSanders* and *@SenSanders*), Hillary Clinton (*@HillaryClinton*), and Donald Trump (*@realDonaldTrump*). Winona LaDuke (*@WinonaLaduke*) is a former politician, tribal activist, and environmentalist. Tara Houska (*@zhaabowekwe*) was the Native American advisor to Bernie Sanders during the time of this study. The Goldman Environmental Prize (*@goldmanprize*) is the world's largest award for recognizing grassroots environmental activists [79]. Finally, Indian Country Today Media Network (*@indiancountry*) is a news media network that provides a platform for Native American journalism and issues [92].

When examining the users mentioned in the general tweets filtered with the keywords[11], we find that the users who are mentioned are much more prevalent in the general data set than in the Native American advocates data set. Moreover, a more substantial collection of the U.S. presidential candidates are represented, including Republican candidate Donald Trump (*@realDonaldTrump*), Senator Ted Cruz (*@tedcruz*), Senator Hillary Clinton (*@HillaryClinton*), Senator Bernie Sanders (*@BernieSanders*), Senator Marco Rubio (*@marcorubio*), and Senator John Kasich (*@JohnKasich*). Similar to the most-mentioned users in the Native American advocates data set, news media is represented, including the mainstream mass media networks (*@FoxNews* and *@CNN*).

[10] "President of the United States."

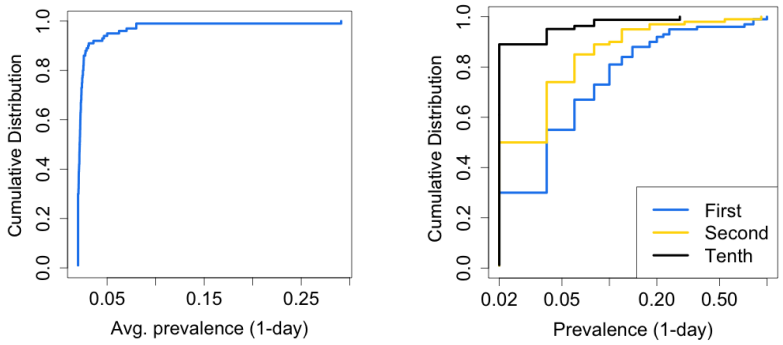
[11] While the majority of the 15 most mentioned user accounts in the general data set referenced U.S. political action, we found that the "vote" keyword captured content that pertained to irrelevant votes and polls (e.g. for Nickelodeon's Kids' Choice Awards for entertainers or Radio Disney's poll for top artists). To ensure that our comparison of mentioned usernames in both data sets makes contextual sense, we discard a total of 6 irrelevant usernames that were captured by our keyword filters (Table 5.6) when applied to the general data set.

5.4 Identifying Sub-communities

To address the second research question, we use clustering methods and sequence analysis to ascertain comprehensive sub-communities (as defined in the Definitions section) based on connections between actors and hashtags in addition to topical sub-communities that exhibit stability over time. We examine these sub-communities across all hashtags in the Native American advocates data set and across political action hashtags identified in the "Political action hashtags" section.

5.4.1 Topical Sub-communities

When characterizing the content prevalent to the Native American advocates data set, we identify the top 100 most circulated hashtags and use a codebook to classify the tags into six topical categories (see Table 5.3). In order to identify how different topical issues might unite through similar bases of support, we identify all the actors tweeting or retweeting posts that contain at least one of the top 100 hashtags. We then construct an egocentric graph wherein the hashtag acts as the ego node and all actors who tweet or retweet a post containing the hashtag act as the neighbor nodes in the graph. We then perform a pairwise comparison between the egocentric graphs associated with each hashtag using the Jaccard similarity index (see Equation 4.3). From these individual ego networks, we create a graph where the nodes represent each of the top 100 hashtags and the edges represent the Jaccard similarity between each of the hashtags. Next, we identify sub-communities present in the graph using the Louvain method, which attempts to partition the graph in such a way that optimizes the *modularity*, or the relative density of edges inside each community as compared to the density of edges between communities [17]. With this technique, we identify 29 sub-communities that exist between the top 100 hashtags with a modularity of 0.81. The median sub-community



(a) (b)

Figure 5.4: Cumulative distribution of (a) the average prevalence of actors associated with the top 100 hashtags and (b) the prevalence of the first-, second-, and tenth-most prevalent actors associated with the top 100 hashtags in the Native American advocates data set.

consists of 3 hashtags ($\sigma = 2.7$) and the median Jaccard similarity present within a community is 0.18 ($\sigma = 0.11$). We provide an overview of the five sub-communities that have the largest average actor bases in Table 5.7.

We next examine the stability of topical sub-communities over time. To do this, we create hashtag-centric ego graphs for each of the top 100 hashtags as they exist in each day of our data set. We examine the presence of each ego's neighbor at each of the time-periods to determine the actor prevalence throughout the entirety of our observation period. As with our analysis of hashtags, actor prevalence indicates the comprehensive degree of churn surrounding each hashtag. Figure 5.4a plots the distribution of the average actor prevalence associated with each of the top 100 hashtags. The average prevalence is 0.022 ($\sigma = 0.029$). *#hiring* has the greatest average actor prevalence with 0.29. We also plot the distributions of the prevalence associated with the first-, second-, and tenth-most prevalent actors associated with each of the top 100 hashtags in Figure 5.4b.

We find that only 26% of the hashtags have an actor that is prevalent for at least 10% of the observation period and only 5% of the hashtags have an actor that is prevalent for

ID	Hashtags	Avg. Jacc. Index	Avg. # actors
1	*#tairp, #freeleonard-peltier, #Indigenous*	0.006	591.17
2	*#mmiw, #idlenomore, #cdnpoli, #turtleisland*	0.009	183.38
3	*#nativelivesmatter, #blacklivesmatter*	0.006	96.25
4	*#facebook, #india, #colonialism*	0.009	61.17
5	*#art, #appropriation, #closethegap, #culture, #lawyers*	0.008	56.10

Table 5.7: Overview of the five largest topical sub-communities in the Native American advocates data set as identified by the Louvain method.

at least 25% of the observation period. The four hashtags that have at least one actor that is prevalent for the majority of the observation period (i.e. more than 24 days) are: *#Indigenous, #ndn, #nativeamerican, and #hiring*. It is also worth noting that 3% of the 2,839 actors involved with *#Indigenous* interacted with the tag on at least two different days between February 11 and March 31.

5.4.2 Political Action Sub-communities

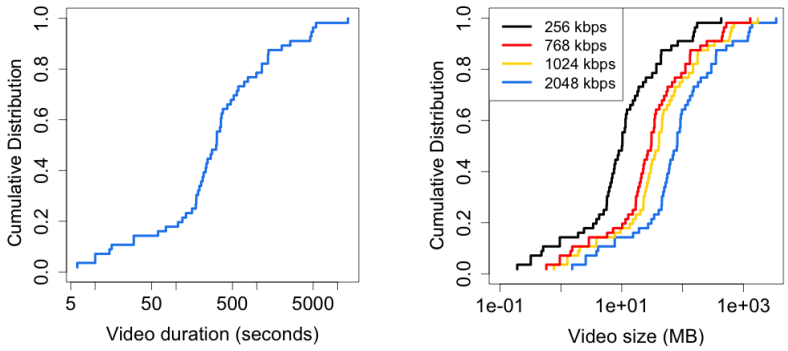
As in the Content Analysis section, we separate the political action hashtags associated with the top 100 hashtags in the Native American advocates data set in order to better understand actor engagement around political action. When examining the sub-communities identified via clustering of hashtag-centric ego graphs, we find one sub-community that contains half of the political action hashtags identified in the Content Analysis section. The hashtags that comprise this community include: *#nativesforbernie*, *#feelthebern*, *#apachestronghold*, *#nativevote*, *#nativevote16*, *#saveoakflat*, and *#israel*. *#saveoakflat* and *#apachestronghold* represent a movement spearheaded by

tribes in Southern Arizona that challenges Congress' right to distribute sacred land to a foreign copper mining company without conducting environmental impact studies or consulting tribes [8]. The topics in this sub-community consist of an average of 28.1 unique actors and the average Jaccard similarity index between hashtags comprising the sub-community is 0.15.

When examining the stability of political action topical sub-communities using the day-long prevalence, we find that political action hashtags exhibit relatively low stability over time. The political action hashtag with the highest level of stability is *#cdnpoli* with the most stable of its 162 actors having a day-long prevalence of 0.22 (mean prevalence is 0.023). One explanation for the large number of one-time actors is the fact that many of the tweets tagged with *#cdnpoli* are also tagged with the hashtag that boasts the most prevalent actors; 31% are co-tagged with *#Indigenous*. In contrast, top political hashtags linked to campaigns (*#nativevote16, #nativevote, #wearebernie, #feelthebern*, and *#nativesforbernie*) have a collective mean prevalence of 0.02, which translates to an actor tweeting/retweeting that hashtag for only a single day in the data set. For these hashtags, the most prevalent actors are associated with *#feelthebern*, which has one actor with a prevalence of 0.04 and the remaining 54 actors have a prevalence of 0.02. Given the relatively low prevalence of these campaigning hashtags, it is noteworthy that on average, only 1.4% of tweets containing them are co-tagged with *#Indigenous*.

5.5 Bandwidth Characteristics

We address RQ3 in light of Indian Country's infrastructural limitations described in the Introduction and Related Works sections. We investigate the impact media richness has on the propagation of individual tweets in the Native American advocates data set. We argue that all tweets are essentially bulletins that enable asynchronous interaction



(a) (b)

Figure 5.5: Cumulative distribution associated with (a) the duration and (b) the sizes of videos embedded in tweets from the Native American advocates data set.

between the poster and audience. However, the richness of individual tweets can vary considerably depending on the presence, type, and size of media embedded in the tweet. Of the 5,172 unique tweets we observe in the Native American advocates data set, 1.7% contain embedded video content, 35.8% contain embedded photo content, and 62.5% do not contain any embedded content. Per Daft and Lengel's definition, we consider tweets with embedded media to be richer than those that lack embedded media [45, 116]. Moreover, we consider tweets with embedded videos or GIFs to be richer than tweets with embedded photos based on the fact that such media offers the "simultaneous transmission of multiple information cues" [116]. Similarly, we consider tweets with embedded videos to be richer than tweets with embedded GIFs, as the audio component lends the expression of a greater "variety of languages" [116].

Overall, we were able to obtain the length of 56 (64%) of the embedded videos; 28 (32%) of the videos corresponded to GIF content that did not have any associated length of time and 3 (3.4%) of the videos were no longer accessible on the Web. In Figure 5.5a, we plot the distribution of the duration of accessible embedded videos. We find that the median duration of embedded videos is 5.3 minutes ($\sigma = 35.4$ minutes). With guidance from the Twitter developer documents, we also report on the sizes of embedded video.

Since Twitter serves various types of devices, the playback rate of video content ranges from 256 kbps to 2048 kbps (depending on screen size and screen orientation)[12]. In Figure 5.5b, we plot the distributions of the sizes of embedded videos as they would correspond to various playback rates. Depending on which playback rate is used, the median video size ranges from 10.1 MB (σ = 68 MB) to 81.2 MB (σ = 543.7 MB). Given the highest bandwidth playback rate of 2048 kbps, only 7.9% of the videos were larger than 1 GB; at the next highest bandwidth playback rate, only 1.3% of the videos were larger than 1 GB. We also use the Twitter developer guidelines to estimate the range of sizes of the 28 GIFs we observe, which defines the minimum GIF duration as 0.5 seconds and the maximum duration as 30 seconds. Given these specifications and the aforementioned playback rates, a GIF can range from 16 KB[13] to 7.7 MB[14]. When examining the sizes of the 1,852 embedded photos, we find that the median photo size is 50.4 KB (σ = 32.9 KB), and the largest observed photo is 269.7 KB. We demonstrate the impact these data sizes have on a hypothetical network infrastructure. Assuming a connection to the Internet that allows for download speeds of 3 Mbps with 100% goodput (which is faster and higher performing than what 85% of Native Americans living on tribal land can access at home [71]), the average video would take between 26.9 and 216.5 seconds to download; the average GIF would take between 42.7 milliseconds and 20.5 seconds to download; and the average photo would take 0.13 seconds to download. Considering that the average length of a Twitter session is 107 seconds [37], waiting for embedded media from a single tweet to download could potentially take a significant portion of the session (if not the entire session).

We next examine the relationship between embedded media and content propagation.

[12]Video Specifications and Recommendations. `https://dev.twitter.com/rest/media/uploading-media#videorecs`

[13]Assuming a 0.5 second GIF with a playback rate of 256 kbps.

[14]Assuming a 30 second GIF with a playback rate of 2048 kbps.

We base our comparisons on tweets from the Native American advocates data set containing embedded content and tweets from the Native American advocates data set that do not contain embedded content using two-sample Kolmogorov-Smirnov tests. Overall, we observe that 66% of tweets with embedded media receive at least one retweet while only 41% of tweets without embedded media are retweeted at least once. Additionally, we find that tweets with embedded media (photo or video) receive higher levels of user engagement ($p < 2 \times 10^{-16}$); on average, tweets with embedded media reach 2.6 users and tweets without embedded media only reach 1.8 users. When examining the prevalence (on a one-day scale) of tweets containing embedded media, we find no significant difference between tweets with and without embedded media; however, we do note that 7 of the top 10 most prevalent tweets contain embedded media. As with the hashtags, we measure churn of specific tweets using the persistence metric at the scale of minutes, hours, days, and weeks. Most tweets do not exhibit persistence at any scale. We find that only 1.8% of all tweets are persistent on the scale of days (i.e., "recurrent") and of these, 66% contain embedded media (of which all but one are photos). Moreover, when analyzing the 0.4% of tweets that are persistent on a week-long scale, we find that 85% contain embedded media.

5.6 Discussion and Conclusion

5.6.1 Issues of Life and Death

Our analysis of Native American political discourse online reveals that the most pressing issues are those with life and death consequences. With respect to "issue-based" hashtags, *#mmiw* and *#nativelivesmatter* garner the largest number of supporters. It is also noteworthy that our analysis of topical sub-communities revealed that three of the

top five topic clusters that garnered the largest number of supporters involved hashtags referring to issues of life and death for Native Americans, including *#mmiw, #nativelives-matter,* and *#freeleonardpeltier*. This is reflective of daily realities for Native American peoples in the U.S. Data collected by the U.S. Department of Justice finds that 34% of Native American and Alaska Native women will be raped or sexually assaulted in their lifetime–more than any other ethnicity group in the country–and on some reservations, Native American women are murdered at a rate 10× the national average [183, 154]. Moreover, a study based on data collected by the CDC found that Native Americans comprised the racial group most likely to be killed by law enforcement [115]. Additionally, violence (including intentional harm, homicide, and suicide) accounts for 75% of deaths for Native American youth between 12 and 20 years old [35]. While these life and death issues loom large in the consciousness of Native American peoples, they are largely absent from the campaigns of major party political campaigns in the U.S. and activists wishing to engage these issues have few outlets in the traditional political sphere.

5.6.2 Mechanisms of Connective Action

While we acknowledge Tufekci's assertion that the topics referenced by the hashtags may be ongoing despite the ephemerality of the hashtag [186], our study of Native American political engagement on Twitter affirms observations made by Bimber and Garret regarding the ephemeral nature of Internet-based political engagement–we observe high churn rates associated with most hashtags, both with respect to occurrence in the data set and with respect to the sub-communities that form around them [16, 74]. In contrast to our general findings of hashtag and sub-community ephemerality, we find that the most enduring hashtag is *#Indigenous*, which was tweeted over 5,345 times by 2,839 users. This hashtag was present in every day of our data set and received some form

of interaction from 85 users multiple days through the course of data collection. These observations confirm assertions made by Tully [187]: that Indigenous solidarity is a political movement (based on cultural identity rather than particular issues, grievances, campaigns, or events) towards self-governance. Moreover, *#Indigenous* was paired with at least one other hashtag in 90% of its occurrences. These observations lead us to believe that Indigenous solidarity hashtags function as a mechanism for connective action between Native American advocates by stitching together a diverse collection of transitory topics for a relatively stable group of actors over time. Thus, the connective action enabled by content dissemination and annotation (i.e. adding hashtags or user mentions to content already circulating) strengthens the voice of Native American advocates and increases momentum for the potential formation of the highly influential Internet-based SMO's described by Bimber [16]. It is important for campaigns to take note of these connective actions and to understand that merely identifying an issue as an Indigenous issue (even if it is also a general issue) can encourage Native American advocates and their followers to connect around it.

5.6.3 Social Media and Infrastructure

Lack of communications infrastructure continues to be a problem for Indian Country that prevents many Native Americans from fully engaging with political discourse that increasingly takes place on media rich platforms [188, 14, 207]. Our results demonstrate that the content that reaches the largest audiences and is the most enduring in Native American advocates' political conversations on Twitter is content that has qualities of greater media richness (i.e., includes embedded media). We find that 66% of the most persistent tweets in the Native American advocates data set contain a photo. Similarly, tweets containing photos receive 24% more retweets than tweets containing video (104%

more retweets than tweets without embedded media). While this finding is consistent with what is observed on Twitter in general [164], investigation into circulation with respect to tweets' persistence and prevalence further highlights the value of embedded photos. Only 1.1% of the most persistent tweets in the Native American advocates data set contain video, whereas 65% of the most persistent tweets contain a photo. Overall, our findings with respect to embedded media agree with Daft and Lengel's assertion that some media is superior to others for communicating information (as measured by propagation and circulation metrics), but it also demonstrates that there are limits to the benefits of increasing media richness, namely the cost of resources required to support richer media might make "less rich" media a more appropriate communication tool. While Native American advocates may not consciously craft and propagate content with bandwidth requirements in mind, the fact that limitations of the underlying IP network may impact information diffusion across the relatively bandwidth-light Twitter platform [186] is worth consideration, particularly if the desired audience for content is connecting from areas with limited ICT infrastructure.

5.6.4 Conclusion

Native Americans represent a politically marginalized group in the U.S., and are also likely to have limited Internet access and connectivity–reducing capacity for political engagement via digital means. We use a post-structural mixed-methods approach to analyze Twitter data culled from influential Native American advocates during the 2016 primary presidential election season. This study reveals that the content propagated by Native American advocates tends to orient around Indigenous solidarity and life-and-death issues for Native American peoples. We find that the most durable sub-communities are those that center on *#Indigenous* and we demonstrate how hashtags that denote Indige-

nous solidarity are the mechanisms through which political connective actions take place between Native American advocates and their followers. Finally, our analysis of Tweets containing embedded media suggests that advocates wishing to propagate content to audiences in Indian Country should enhance communications by embedding small photos rather than larger media files to ensure that richness of communication is balanced with consideration for infrastructural limitations.

5.7 Acknowledgements

This work was done in collaboration with Marisa Duarte, Nicholet Deschine Parkhurst, and Elizabeth Belding. Special thanks are due to Ellen Zegura for her early feedback on data collection methodologies. This work was funded by NSF Graduate Research Fellowship Program under Grant No. DGE-1144085 and the NSF Network Science and Engineering (NetSE) Award CNS- 1064821.

Part II

Network Innovations for Challenged Environments

Chapter 6

Repurposing FM Radio for Content Delivery

In order to bridge the information gap, many Native Americans living on reservations rely on FM radio to receive culturally relevant programming, emergency notifications, local news, and information on civic participation, health, and economics [6, 2, 174]. While FM radio operates at frequencies that have excellent penetration rates for distributing content over long distances and can be received by end users at little to no cost, it is not supportive of real-time, interactive, media-rich content that is pervasive on the Web today.

Studies of Indigenous broadband usage indicate that when Internet services are available, online social networks (OSNs) account for the most frequently accessed sites [31, 129, 193, 202]. While users cite reasons for using social media that are similar to the general population [176, 193], there is also a distinct emphasis on the use of OSNs for building cultural resilience through local social connections, staying connected to geographically distant friends and family [31, 129], and receiving news from alternative media sources [6, 84].

In this work, we propose and evaluate a content dissemination system that leverages the data subcarrier of FM radio transmissions (RBDS) in order to broadcast digital content to co-located communities of users who do not have access to point-to-point communications infrastructure due to issues of affordability or availability (or both). Based on the reported information needs of Indigenous communities [6, 58, 31] and findings from previous studies of Indigenous Web usage [31, 129], we find that content from online social networks (OSNs) is critical to Indigenous efforts towards cultural revitalization and building community connections. Thus, our proposed system focuses on the efficient dissemination of OSN content in particular.

In order to assess the feasibility of OSN content delivery over RBDS, we use a trace-driven approach using a month's worth of data from the `TDV Instagram-2014` data set. In assessing the design and feasibility of our proposed system, we are guided by three research questions:

> **RQ1** *How does a set of co-located users interact with OSN content and what implications does this activity have on bandwidth requirements?*
>
> **RQ2** *How much overlap exists between the social networks of users who are from the same geographic community?*
>
> **RQ3** *Can we leverage social network relationships to reduce overall bandwidth usage while still providing adequate content coverage?*

We address the first two research questions in Chapter 4.2, where we find that Instagram users in the TDV network have a high locality of interest with respect to social media content, only interact with a small percentage of the social media content available to them, and have very limited overlap in their social networks. Given these observations, we approach RQ3 by proposing a content distribution system that would provide access

to OSN content without requiring access to pervasive broadband or cellular infrastructure. We then evaluate the feasibility of such a system using the `TDV Instagram-2014` data set. We demonstrate that it is possible to take advantage of social connections between co-located users in order to reduce high bandwidth content so that *users can receive the most relevant content via technology that is already ubiquitously available.*

6.1 OSN Over RBDS

Leveraging the digital transfer capabilities of FM frequencies has not been explored in the context of bridging the digital divide. The FM radio broadcast data system (RBDS)[1] is a communications protocol that enables FM broadcast stations to embed small amounts of digital information into conventional FM radio broadcasts. RBDS functions as a subcarrier on the main FM radio transmission–its most familiar application is to carry digital information about current FM broadcasts (e.g. song title, artist name, station name) to display to listeners. While it operates over extremely low bandwidth (1.1875 kbps), it has many properties that are appealing for data transmission to areas lacking Internet infrastructure, including: large geographic coverage footprints [161], robustness to error [161], and operational infrastructure [36, 2]. While the bandwidth limitations of the protocol are prohibitive to certain types of data services (e.g. streaming media, interactive applications, real-time services), RBDS does have the potential to provide access to OSN content to areas where FM broadcast stations function as the sole arbiters of media content. This work is the first to characterize the feasibility of RBDS to provide data services to disconnected communities.

It is important to note that use of RBDS to deliver information from OSNs could be implemented in many ways, given the availability of FM receivers. In order to simplify

[1]Radio Broadcast Data System in the U.S. and Radio Data System (RDS) internationally.

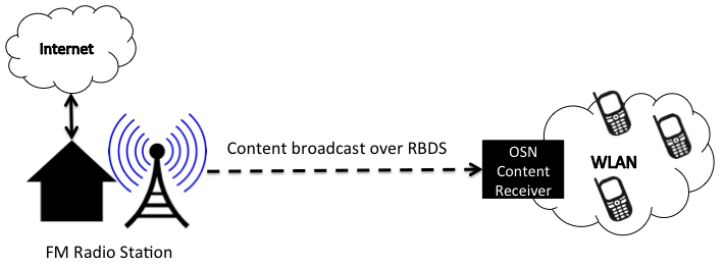


Figure 6.1: Proposed system for the propagation of OSN content over RBDS.

our approach to understanding the feasibility of RBDS for OSN content delivery, we assume the architecture shown in Figure 6.1. Such an architecture would be able to actively download public OSN content and then use a protocol such as RDS-Link [161] to broadcast data to remote, offline repositories where users could access content over a local wireless network.

Our proposed system focuses on delivering OSN content to rural Indigenous users over the FM radio broadcast data system. The focus on OSN content is motivated by studies of Indigenous ICT usage which have shown OSNs to be major contributors to Indigenous cultural revitalization and resilience. Prior surveys have revealed that social networking sites are the most popular means of everyday communication in the Sioux Lookout community [31, 129, 76]. Moreover, these studies have also found that OSNs play a potentially powerful role in Indigenous cultural preservation and revitalization. Many surveyed users reported using OSNs to post information about upcoming cultural events [31, 129], to find information about cultural events and look at photos and videos of the events [31], to look at pictures of family and ancestral land [76], to provide and seek support for patients participating in Indigenous telemedicine programs [31], to participate in political and social movements advocating for Indigenous rights and justice [6, 125], and to communicate in their native language [31]. The most popular OSN may differ from community to community; in some of the studied Indigenous communities Instagram

is the most popular OSN, in others it is Facebook, and in still others, it is MyKnet[2]. We use Instagram to investigate the feasibility of delivering OSN content over RBDS and we believe that the general principles will hold for the delivery of content from any OSN. More important than the specific OSN is providing access to OSN content and the benefits it can provide to Indigenous communities. Our OSN content delivery system over RBDS enables more people living in rural and remote areas without broadband coverage to gain access to this beneficial content *using a technology and infrastructure that is already ubiquitous: FM radio.*

We also note that our proposed system is not limited to delivering OSN content. In rural areas, it is still difficult to make reliable emergency announcements without ubiquitous cellular or telephone coverage. In cases of severe weather, natural disasters, and water contamination, our proposed system could provide information about safety measures, evacuation plans, and access to alternative resources.

6.2 Content-Oriented Broadcasting

Approach	Median coverage per user %	Mean coverage per user %	Std. dev. coverage per user %	Jain fairness index
Random	33.1	33.8	9.5	0.88
FCFS	32.3	33.3	9.2	0.88
P-Coverage	52.2	53.6	18.5	0.86
P-CP	77.8	65.5	30.1	0.77
Round robin	54.3	55.0	14.5	0.89

Table 6.1: Overview of the performance of scheduling approaches.

The analysis in Chapter 4.2 has shown that social connections between more geo-

[2]A First Nations owned OSN based in northern Ontario.

graphically proximate users are stronger and that stimulated content tends to be viewed by a higher number of co-located users. Thus, we address RQ3 by evaluating different approaches to scheduling OSN content over RBDS. In this section, we compare and evaluate five different scheduling approaches: random scheduling, first come first serve (FCFS), coverage-based priority scheduling (P-Coverage), cluster-based priority scheduling (P-CP), and round robin scheduling. The software used to generate the resulting scheduled content is available at `http://github.com/mvigil90/RBDS_Scheduling`. While all of the approaches investigated in this chapter have been evaluated in the context of networks, these prior works assume a one-to-one connection between one content provider and one receiver. Prior to our evaluation context, there has not been a data-based evaluation of how these approaches perform in a broadcast scenario where broadcast content can be relevant to multiple receivers.

6.2.1 Description of Scheduling Approaches

We provide an illustration of the general scheduling approach in Figure 6.2. For all scheduling approaches, we assume that there is sufficient storage at the FM radio station to store all incoming content prior to scheduling it for broadcast. This collection of content includes all content that was published during the current scheduling period. The amount of content that is broadcast per scheduling period is determined by the length of the scheduling period using

$$W = T \times B \tag{6.1}$$

where W is the size of the broadcast window, T is the length of the scheduling period, and B is the bandwidth capacity. For our evaluations, we use a scheduling period of 3600 seconds (one hour). This means that we schedule enough content to fill one hour's worth

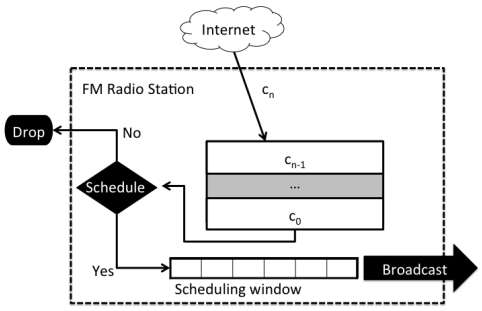


Figure 6.2: General scheduling approach, where published content objects, c, are identified by a unique identifier and placed in an array until the end of the scheduling period when content is scheduled for broadcast.

of broadcasts and drop the remaining unscheduled content in order to provide timely delivery of transmitted content. We assume the RBDS transmission operates without error, yielding a net bandwidth capacity of 1.1875 kbps. This leads to a broadcast window of 534 KB. Because video content objects are over 10× larger than image content and only account for 6.4% of the overall content objects that occur in our evaluation set, we only evaluate scheduling approaches based on how they schedule image content. Since we assume that no transmission errors occur, all content that is scheduled into the broadcast window is actually broadcast.

Random. To provide a standard against which to evaluate the other approaches, we schedule content using a random approach. For each scheduling period, the random scheduling approach selects a random content object from the queue of published content and then either schedules it or drops it with equal probability. Once the scheduling window is full for the broadcast period, any remaining content is dropped.

FCFS. We evaluate a naive first come, first serve approach to schedule data for broadcast in each scheduling window. Our implementation of the FCFS approach fills the scheduling window with the most recent content every scheduling period, then drops the remaining unscheduled content. In the context of broadcast over RBDS, large content flows would correspond to users following prolific content creators, as well as users who

follow many content creators.

Coverage-based Priority. Priority scheduling is similar to the FCFS approach, but instead of scheduling content based on the time it was published, we prioritize content based on the coverage provided by the creator of the content. The more followers within a set of co-located users that a content creator has, the higher its content is prioritized. We call this priority-coverage (P-Coverage). We expect this approach to bias in favor of users who follow content creators with high coverage values. Conversely, we expect priority-coverage to bias against users who follow more obscure content creators.

Cluster-based Priority. The cluster-based priority scheduling approach uses the clustering potential (CP) associated with content creators in order to prioritize content. We find the clustering potential associated with a content creator, c, by calculating the summated clustering coefficient [209] of each user who would receive a given content object using the following equation:

$$CP_c(x_0, x_1, ..., x_n) = \sum_{i=0}^{n} \frac{T_{x_i}}{A_{x_i}} \tag{6.2}$$

where x_i represents a user that follows content creator c , n is the total number of users that follow content creator c, T_{x_i} is the number of creators that form a triadic closure with x_i, and A_{x_i} is the number of creators in x_i's follow network. In the context of the TDV Instagram social network, a user's clustering coefficient determines how embedded a user is within the TDV network. While triadic closures, or three nodes connecting to each other, provide some clue as to how embedded a user is within their local network, the clustering coefficient normalizes this value so that relative embeddedness can be compared across different subgraphs of the network. Given the publish-subscribe nature of the Instagram platform, we consider single-directional links to count towards a triadic closure. The clustering coefficient in the context of co-located users provides an indication

of how many users might find information destined for a particular user to be relevant. The approach that prioritizes based on content creators' CP values is called priority-CP (P-CP). Priority-CP is expected to bias in favor of highly embedded users and users with smaller follow networks as it is easier to have a high clustering coefficient with a smaller overall follow network. We expect both priority approaches to result in starvation of users who are dissimilar from the majority of co-located users with respect to content interests and users who are not deeply embedded in their follow network.

Round Robin. In order to ensure that each of the co-located users receives some portion of their followed content, we evaluate a round robin approach to content scheduling. For the round robin approach, we create a queue for each user. As content is published, if a user follows that content creator, the content is added to the user's queue. Content in user queues is ordered based on the community coverage provided by the content creator. The round robin approach cycles through the user queues at the end of each scheduling period. While there is room available in the scheduling window, it adds the content at the front of each user queue to the scheduling window. When a content object is added to the scheduling window, it is removed from the queue of any other user that has the object in their queue. This would occur when multiple users are following the same content creator. Round robin provides an equal portion of bandwidth to each user. Some users opportunistically benefit from other users' access to bandwidth because they follow the same content. Since some users follow significantly more content than others, round robin is expected to bias against users that require an overall larger portion of the bandwidth resources.

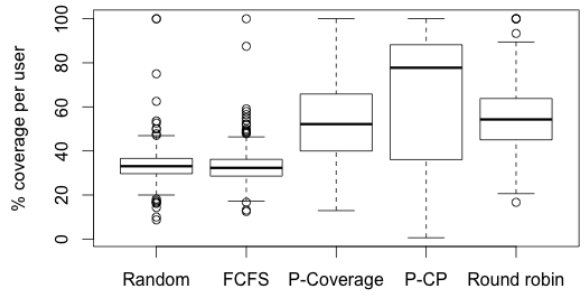


Figure 6.3: The proportion of content coverage provided to each user.

6.2.2 Performance Evaluation

We evaluate each of the scheduling approaches by scheduling one representative[3] ($p < 2.2 \times 10^{-16}$) month of the data set we collected, with respect to the number of content objects posted by creators and the distribution of coverage values associated with posted content objects. We use a one hour scheduling window for each approach. Based on the specifications of RBDS, this window size enables a maximum of 534 KB to be broadcast per hour. Our prior study of the TDV network revealed that the median size of downloaded Instagram image content was 20 KB [202]. As no file size is provided as part of the content meta data we collected, we make the simplifying assumption that all image content is 20 KB. An overview of each approach's performance is provided in Table 6.1, which reports on the distribution of the percentage of followed content that was received per user and the fairness of the approach with respect to the coverage provided to each user.

Completeness. We examine completeness from the perspective of individual users and the group of users. We measure the total coverage provided by a particular scheduling approach by identifying the number of OSN content objects that the approach

[3]Based on the Two-sample Kolmogorov-Smirnov test comparing the distribution of coverage values and the number of content objects posted per month.

scheduled over our evaluation month relative to the total number of OSN content objects that were published during that month. Each scheduling approach was able to provide a total coverage of 35.5%. We measure the coverage per user by identifying the total number of content objects that would be received by each user during the evaluation month, then identifying what percentage of that content was actually broadcast using each scheduling approach. The distributions of the coverage per user provided by each approach are plotted in Figure 6.3. On average, individual users received the highest percentage of their followed content when using the round robin scheduling approach. These results are unsurprising for several reasons. By scheduling content based on user demand in a fair way, the round robin scheduling approach is able to address the content needs of individual users. Because several content creators are followed by more than one user in the TDV network, the round robin scheduling approach was able to opportunistically satisfy the needs of several users when scheduling content based on the needs of a single user.

The FCFS scheduling approach performed the worst with respect to coverage per user. Given its prioritization based on publication time, it is unsurprising that FCFS would not provide the optimal approach for scheduling bandwidth in an overloaded environment. FCFS is unable to discern community value associated with content because it operates without any concept of relative content demand. This causes users who follow highly productive content creators, users who follow many content creators, and users who follow more video content to receive a much larger portion of the available bandwidth.

We also note that priority-CP had a high coverage rate with respect to the individual, but varied significantly between users and ultimately provided the most unfair division of resources (we discuss this further in the section "Fairness"). Priority-CP tends to favor content arriving for small, tightly knit sub-groups of users and leads to starvation of less embedded users. While it was not the best scheduling approach for the Instagram

context, it could be useful in other scenarios. For instance, if broadcast were the first part of an information relay, targeting content towards a clustered group of individuals may lead to more effective information dissemination throughout the community.

Fairness. In order to evaluate the fairness of each scheduling approach, we calculate the Jain fairness index [104] over the number of content objects each user would receive. The Jain fairness index is calculated as follows:

$$jain(x_0, x_1, ..., x_n) = \frac{(\sum_{i=0}^{n} x_i)^2}{n \cdot \sum_{i=0}^{n} x_i^2} \tag{6.3}$$

where x_i represents the percentage of user i's content requests that were broadcast (coverage per user), given n co-located users to receive scheduled broadcasts.

The Jain fairness index for each scheduling approach is reported in Table 6.1. It is expected that the round robin approach would be the fairest of those evaluated on the data, as it is an inherently fair approach because it provides equal opportunities for all users to receive content that is relevant to them. Because of the broadcast nature of RBDS and the fact that multiple co-located users can follow content published by a single content creator, scheduling content from a single user's queue could lead to multiple users receiving content outside of what would be scheduled from their own queues. This causes co-located users with overlapping follow networks to have a slight advantage over users who do not follow networks that are similar to other co-located users. The least fair approach is priority-CP. This is expected since prioritizing content destined for a highly clustered set of users would starve out a significant portion of content given the lack of observed overlap and clustering between TDV users (the mean clustering coefficient is 0.039). We also find that FCFS is less fair than the random scheduling algorithm, which is unsurprising given the inherent unfairness of FCFS, which favors users who follow highly productive creators and tends to starve out smaller flows. Even though the

random approach and FCFS provide significantly lower coverage per user, they still have high fairness values. These approaches are fair despite the lower coverage values because the coverage they provide has very low variance. The variance in coverage per user is 0.7% for the random approach and 0.8% for FCFS. Round robin provides a higher level of fairness than both priority-coverage and priority-CP because round robin provides a much lower variance in coverage per user. The variance for the round robin approach is 2.1%, the variance for priority-coverage is 3.4%, and the variance for priority-CP is 9.1%.

Timeliness. By only scheduling content that arrives during the current scheduling period, the delay between the time content was published and the time it is broadcast is always under one hour[4]. While this delay is reasonable given our observations of users' responsiveness to content in the section "Publication time," we note that if we were to relax this delay requirement, certain approaches would be able to provide a higher level of coverage to individual users. We test the impact of relaxing delay constraints by storing unscheduled data for future scheduling periods (instead of dropping unscheduled content as described in the beginning of the section "Description of approaches"). This means that scheduling approaches would be able to fill the scheduling window with content published during the current scheduling period and unscheduled content from previous a scheduling period. We then evaluate performance over the same representative month of content for approaches that could increase coverage by scheduling stored data: priority-coverage, priority-CP, and round robin.

We plot the distributions of delay incurred by scheduling stored content in Figure 6.4. Although we expected the delay to increase for all approaches that schedule stored content, priority-CP was able to maintain an average delay time below an hour, with a median delay of 46 minutes and a standard deviation of 11.5 hours. Since the scheduling window is limited to 534 KB per hour, only 26 content objects with the highest CP values

[4]Only scheduled content has associated delay times; unscheduled content is not considered when calculating delay.

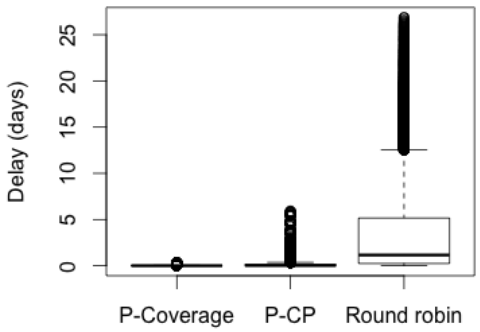


Figure 6.4: The distribution of delays incurred by scheduling stored content.

are broadcast. When we examine the distribution of the CP values associated with incoming content across time, we find that on average, 36 of the content objects published per hour have a CP value greater than the median CP ($CP = 0.01558$) value and 18 of the content objects published per hour have a CP value greater than the third quartile CP value ($CP = 0.05556$). This means that each hour, content that is broadcast is likely to have been published during the most recent hour, leading to generally low delays. The median delay for priority-coverage is 4.9 days (standard deviation = 6.1 days) and the median delay for round robin is 1.2 days (standard deviation = 5.0 days).

Despite the longer delay times associated with broadcast content, we do find that the coverage provided by the approaches increases. Figure 6.5 plots the distribution of coverage per user provided by each approach scheduling stored content. Overall, each approach is able to provide a total coverage of 36.4%. This is an increase compared to the approaches that schedule only content published in the current scheduling period because some hours of the month do not have enough content publications to fill an entire scheduling window. With the ability to schedule content stored from previous scheduling periods, the approaches are able to fill the entire scheduling window during every

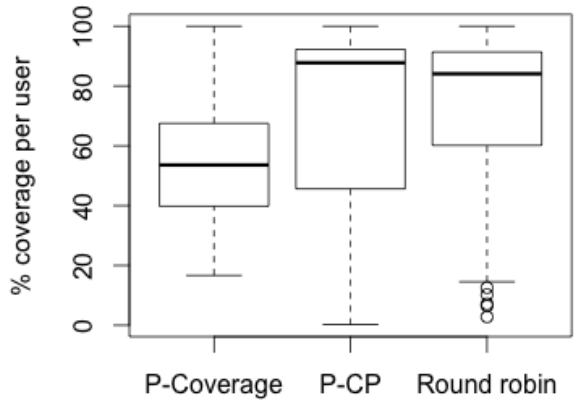


Figure 6.5: The distribution of content coverage provided by approaches using stored content.

scheduling period because there is always a backlog of unscheduled content to fill any gaps. In addition, the round robin approach and priority-CP approach are able to increase the coverage per user (the coverage per user provided by priority-coverage remains within 1% of what was provided for scheduling without stored content). The median coverage per user for priority-CP is 87.8% (standard deviation=32.1%) and the median coverage per user for round robin is 84.1% (standard deviation=23.8%). Consequently, the Jain fairness index changes for these two approaches, and when operating over stored content, the Jain fairness index for the priority-CP approach decreases to 0.75 and the Jain fairness index for the round robin approach increases to 0.91.

There is a clear trade-off between timeliness and coverage. Storing unscheduled data for future scheduling periods ensures that content has more opportunities to be broadcast. However, each content object has to compete against a growing reservoir of content in order to be scheduled. This can lead to longer delays between publication time and broadcast time. One way this trade-off can be negotiated is by regularly flushing un-

scheduled content after it has been stored for longer than a certain period of time. The flush period would be based on the delay requirements and the average volume of data published per hour. Another way to balance delay and coverage would be to schedule only content published during the scheduling period and at certain times, consider stored content for broadcast. Our analysis in the section "Publication time" reveals that Instagram users have an eight hour trough in activity, meaning they are not likely to be checking their phone for incoming posts during this period. During the hours of the day that users are checking content more frequently, it would make sense to only broadcast the most current published content. Conversely, the period of inactivity might be an opportunity to broadcast stored content.

6.3 Discussion and Conclusion

Here we discuss some of the issues of practical information dissemination via social RBDS, including user interaction, moving beyond social content, and bootstrapping. Although this chapter has focused on a Native American community, it is important to note that the proposed social broadcast over RBDS could be used beyond the reservation context in developing areas where there is limited broadband coverage or where communication services are prohibitively expensive.

6.3.1 Generalizability

The data used to evaluate our content delivery system were generated by Instgaram users from six different Native American reservations in San Diego County. While the OSN usage data that we characterize in our analysis are specific to TDV Instagram users, research exploring the social media practices of both Indigenous and rural communities has revealed similar characteristics of high locality of interest with respect to content and

social connections [129, 106, 107, 216, 77]. Here we discuss how the observed locality of interest in other communities might be leveraged by our proposed OSN content delivery system over RBDS. We provide two examples of how our findings would transfer to other communities based on their OSN usage patterns. The first example describes OSN usage in the Sioux Lookout region, which is home to Indigenous Ojibway, Oji-Cree, and Cree peoples. The second example describes OSN usage in a rural village in Africa.

Survey-based studies performed by Molyneaux et al. on the social networking practices of Indigenous peoples in the Sioux Lookout region of Northwestern Ontario reveal that OSNs are the most frequently used Web applications and that members of the community use technology to help preserve their Indigenous culture by posting cultural content to OSNs, posting announcements about upcoming cultural events, and reading and listening to other OSN users' culturally relevant posts [129]. In this study, 60.2% of the 588 surveyed users accessed OSNs to daily interact with co-located OSN users. This is similar to our finding that TDV Instagram users interacted with content from co-located users 46.6× more often than content generated by users outside the TDV network. The locality of interest suggests that OSN content transmitted over RBDS can provide high coverage to this community of users, particularly if content pertains to local cultural topics.

Looking to increasingly global populations, locality of interest (particularly in small, rural communities) is strong. Previous studies of network traffic locality of a rural village in Macha, Zambia reveal that users are highly involved with online social networks and that the majority of Facebook instant message exchanges are between co-located users [107]. Further study of OSN usage in the Macha community revealed that 46% of downloaded Facebook images were not unique images [107] and the top 10 monthly downloaded Youtube videos received an average of 541 views per month with a standard deviation of 2,580 views per month [216]. This implies that the Macha community could

be well-served by our proposed OSN-to-RBDS system given the high rate of overlapping content interests exhibited by co-located users.

6.3.2 Interaction and Feedback

The scheduling approaches we evaluated had access to users' follow networks prior to scheduling content. Since follow networks are dynamic, there would need to be a mechanism that allows users to update their social network information in order to receive content from all of the content creators they follow. Additionally, the data sets used to evaluate the feasibility of broadcasting OSN content via RBDS did not incorporate information on relevance based on ratings or preferences–from the TDV Instagram users as a group or as individuals. If users had access to a text-message based interface, they would be able to assign priority preferences to different types of content or to provide feedback on the relevance of content that was broadcast. With this integrated information, it would be possible to develop models that predict relevance of incoming content to a specific user and to a group of users.

6.3.3 Access

In order to deploy a system like the one described in Figure 6.1, there are a few bootstrapping concerns that must be addressed. If users wish to receive OSN content over RBDS, there needs to be an interface for them to grant access to their social network information or subscribe to certain content. While this is possible with the current APIs for most OSN platforms (e.g. Instagram, Facebook, Twitter) and sites that publish RSS feeds, it still requires users to have some form of Internet access over which to grant these permissions or requests. With federal grant programs focusing on providing broadband access to schools and libraries on reservations, there may be public Internet access avail-

able in the community that would allow users to provide initial information to an RBDS content broadcast system. However, on reservations that span a significant geographic area, the nearest Internet connection may be tens of miles away, and access may require significant time and fuel resources even if the Internet service itself is free [178]. One way to address this challenge involves a text-message based application that could interface with a service that grants access to a users' social network information on their behalf. Although high speed mobile broadband services such as 3G, 4G, and LTE are not pervasive in Tribal lands, 2G services (which work predominantly for voice and text messaging) have significant penetration and most service plans include unlimited text-messaging or inexpensive "voice and text-only" plans.

6.3.4 Conclusion

While it is easy to imagine that issues of the digital divide are endemic to developing nations, there are still communities in developed nations that do not have access to the Internet due to lack of availability. In this study, we focus predominantly on the digital divide present on Native American reservations in the U.S. [70]. Prior studies investigating the informational needs of reservation communities have emphasized the importance of OSN content for connecting users together through Indigenous language [129, 76], cultural practice [31], experience sharing [31], participation in political and social movements [6, 125], and daily communication [129]. Understanding the information habits of Indigenous communities enables researchers to develop new solutions that help users that live beyond the borders of Internet service connect to each other through existing platforms for computer-supported collaboration (e.g. OSNs). Our characterization of OSN usage by TDV Instagram users reveals a high locality of interest that is consistent with reports of OSN usage in both Indigenous [129, 31, 76] and rural [107, 77]

communities around the world. Based on these observations, we propose a system that disseminates OSN content via RBDS. We show that RBDS holds unlocked potential for content dissemination in communities that do not have Internet access, and our evaluation of various content scheduling approaches suggests that RBDS bandwidth can be shared fairly amongst co-located users even when they have relatively dissimilar content interests. Through the implementation of our system, users who have been historically excluded from the unique collaboration opportunities afforded by OSNs could participate using infrastructure and technologies that are already ubiquitous and affordable.

6.4 Acknowledgements

This work was done in collaboration with Elizabeth Belding and Matthew Rantanen. We would like to thank Joseph Perralta and Geoff Herrin of the Southern California Tribal Digital Village Network for their help with the collection of data used in this evaluation of this system. This work was funded by NSF Graduate Research Fellowship Program under Grant No. DGE-1144085 and NSF Network Science and Engineering (NetSE) Award CNS-1064821.

Chapter 7

FiDO: Content Delivery for Challenged Network Environments

Often, people living in these disconnected communities rely on Internet hot-spots (WiFi 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) [64, 157, 13, 146, 153] and Internet cafés [39, 40, 82]. 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* [39, 82, 153]. 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 [39].

Browsing activity is distinct 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 and content [177]. 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 [119, 9, 40] and Web content recommender systems [152] 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 individual for which the system browses the Web. 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 [202, 199] we ask the following research questions:

RQ1: *How much of a household's Web browsing needs can be met opportunisti-*

cally?

RQ2: *To what degree can the members of a disconnected household rely on their surrounding community to identify relevant and interesting content on their behalf?*

Motivated by our observations of locality of interest within the tribal networks [202, 199] and the combined potential of opportunistic networking and recommender systems, 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 then 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 [101], 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.

7.1 System 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

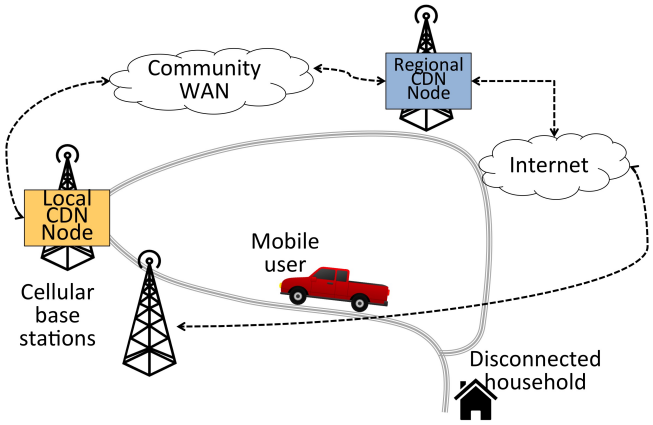


Figure 7.1: An example of FiDO's operation, where a member of a disconnected household opportunistically collects relevant content from CDN nodes located on cellular base stations along their commute.

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 7.1. 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 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 7.2, 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 7.2.2 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

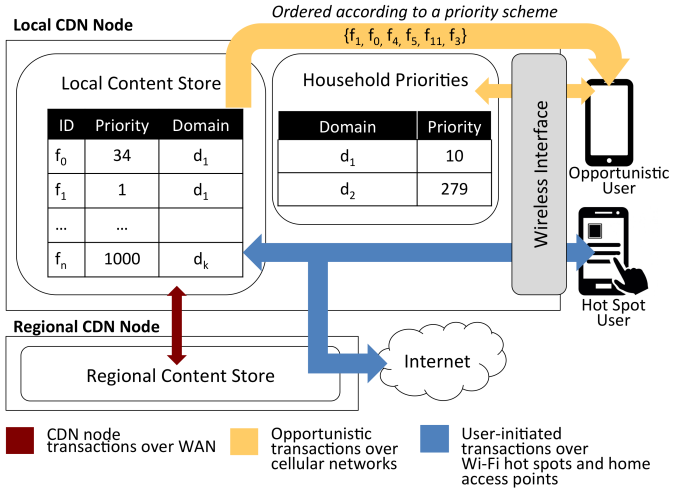


Figure 7.2: FiDO data flow diagram. Arrows represent the flow of content. Content is browsed by users connected to the Internet at home or at WiFi hot spots. FiDO fetches and stores content (which has been filtered using the browsing patterns of the surrounding community) on behalf of disconnected households. When a user from a disconnected household connects opportunistically, FiDO pushes content to the user's device according to a prioritization scheme.

discussed in Section 6.3.

7.2 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 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.

7.2.1 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 7.1, where a user collecting content on behalf of

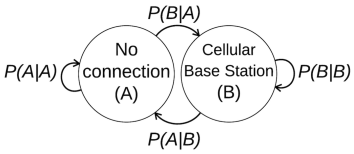


Figure 7.3: Connectivity state machine used in simulation.

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)
$P(B \mid A) = 0$ $P(A \mid B) = 1$ $P(B \mid B) = 1 - P(B \mid A)$ $P(A \mid A) = 1 - P(A \mid B)$	$P(B \mid A) = \begin{cases} 1, & t < c_w \\ 0, & t \geq c_w \end{cases}$ $P(A \mid B) = \begin{cases} 0, & t < c_w \\ 1, & t \geq c_w \end{cases}$ $P(B \mid B) = 1 - P(B \mid A)$ $P(A \mid A) = 1 - P(A \mid B)$

Figure 7.4: State machine transition probabilities based on the time of day.

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 7.3) 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 users' 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[1] per day is 55.87 minutes [51]. We assume that this average number of driving minutes occurs at the average rural speed limit of 75 miles per hour [32] 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

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

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.) [51]. 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 7.4 presents the transition probabilities based on the time of day. The one-way commute time, c_w, is a modeled from a normal distribution with a mean of 27.94 minutes and a standard deviation 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.

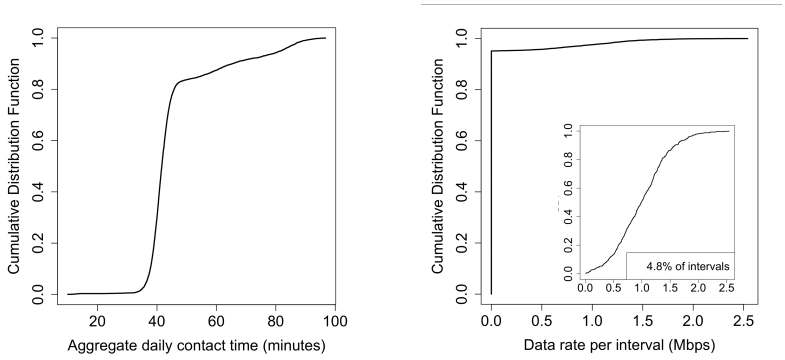
In order to simulate access restrictions associated with specific Web files, users can only receive a Web file if they have actually received it in the actual traces of use. Our simulations assume that members of a household can entrust their access credentials to the user who is connecting to FiDO on their behalf. The significance and complications of these assumptions are discussed in Section 6.3.

7.2.2 Prioritization Schemes

At the most general level, the system outlined in Figure 7.1 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.

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. 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.

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



(a) (b)

Figure 7.5: 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.

the files downloaded from that domain based on the number of times the community has downloaded the file.

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

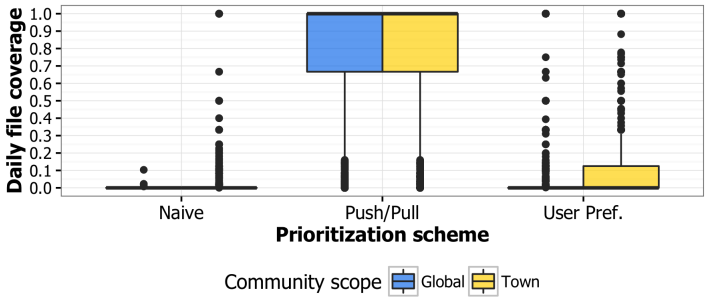


Figure 7.6: Boxplot of distributions associated with daily file coverage provided by each of the prioritization schemes assuming an average total daily commute time of 55.87 minutes traveled at 75 miles per hour. We graph the coverage provided when using recommendations by the household's town in yellow and recommendations by the global network community in blue.

our first research objective: *How much of a household's Web browsing needs can be met opportunistically?*

7.2.3 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 7.2.2. 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 Chapter 3.3). 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 Figure 7.5, 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 of 0.99 Mbps ($\sigma = 0.24$ Mbps).

We plot the distributions of file coverage achieved by each prioritization scheme in Figure 7.6. 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.

7.2.4 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 7.2.3 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 7.2.2. 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 [127, 24, 42]. 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 [24, 42]. In a more recent study of Web graph structure, Muesel et al. observe that the average out-degree for a

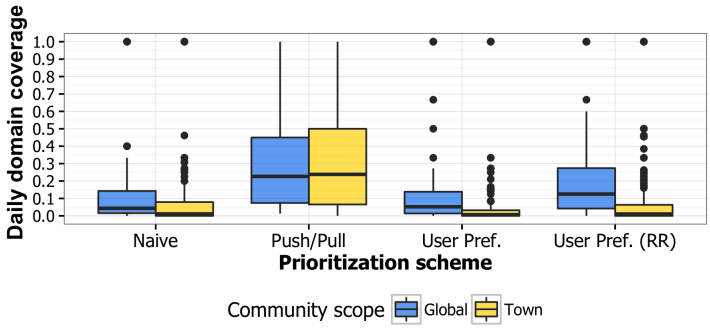


Figure 7.7: Distribution of the daily coverage of Web domains expected by household members at the end of each day of the simulation.

Web page is 36.7 and the tail of the distribution decays at an exponent rate of 2.77 [127]. 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 $\alpha = 2$) [91]. 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 3.3), 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 7.7, 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.

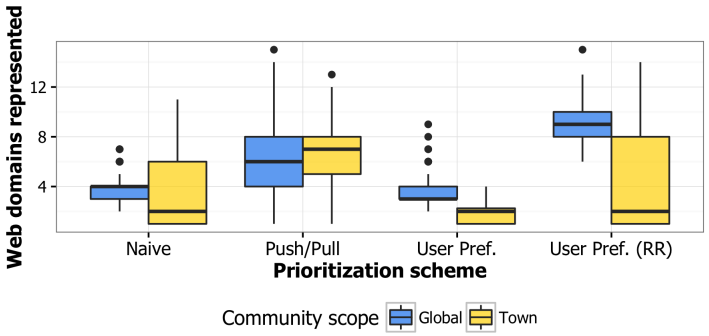


Figure 7.8: Distribution of the number of Web domains presented to household members at the end of each day of the simulation.

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 3.3. 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 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

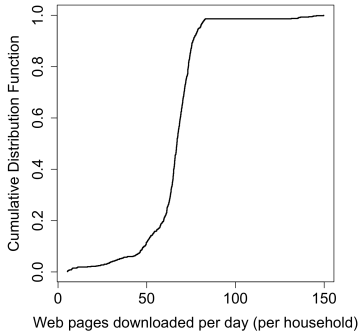


Figure 7.9: Distribution of the number of Web pages provided to each household at the end of each day.

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 7.8). 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 Figure 7.9, 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 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

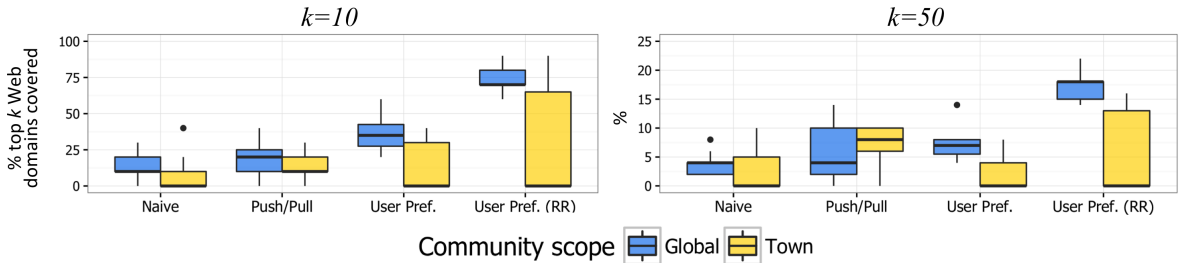


Figure 7.10: Distribution of the percentage of the top k Web domains accessed by each household that are covered by FiDO using each prioritization scheme.

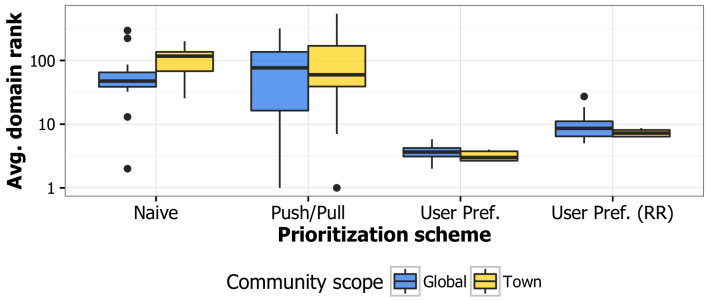


Figure 7.11: Distribution of the average rank associated with Web domains that have pages pushed to users. Lower rank is better.

domains[2] 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 7.10. 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 7.11. 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

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

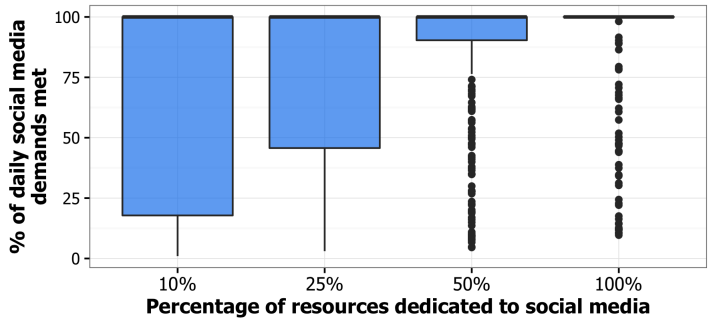


Figure 7.12: Distribution of the percentage of households' daily social media demands that are met using the hybrid prioritization scheme.

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.

Hybridized prioritization. Related work [129, 208, 76, 31] as well as our own previous work [204, 202], 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 4.1.1). 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

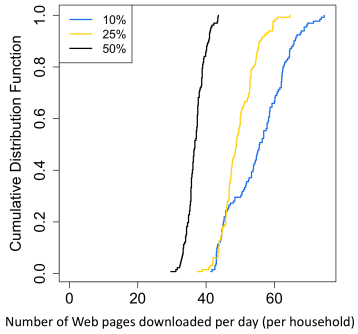


Figure 7.13: Distribution of the number of Web pages downloaded per day per household for the hybrid prioritization scheme operating with 10%, 25%, and 50% of resources dedicated to downloading social media content only.

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 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) are 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 7.12, 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 7.13), 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.

7.3 Discussion and Conclusion

While we focus our work on tribal reservations due to our current partnerships, our work is more broadly applicable to rural communities in general. Systems operation in rural, disconnected communities is non-trivial [21]. 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 [190]. 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 [145, 54]. 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, it is quite feasible that 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 con-

tent 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 [123]. 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 [123, 192, 185]. 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. Thus, 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.

7.3.1 Conclusion

Web access is still far from ubiquitous and even in developed countries, pernicious digital divides persist [99, 192, 58, 6]. 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

chapter, 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.

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.

7.4 Acknowledgements

This work was funded by the NSF Graduate Research Fellowship Program under Grant No. DGE-1144085NSF and grants NSF-1563436 and NSF-1637265. We would like to thank Valerie Fast Horse, Tom Jones, and Justin Hall at Red Spectrum Commu-

nications for their assistance in the collection of data used in this work.

Chapter 8

Applications for Challenged Environments

Due to lack of reliable, high-bandwidth connectivity, users in poorly connected communities are often prohibited from fully participating in important discussions and activities that take place over the Internet. Indeed, not only do they miss out on opportunities to engage with global organizations and movements, but they also miss out on opportunities to engage locally in a virtual space. A particularly salient example is from an analysis of Internet traffic generated by users in a poorly connected Zambian town called Macha. Social networking via Facebook was one of the most pervasively used applications in the network and it provided a platform for users to share content (e.g., photos, videos, music files) they had created and to congregate and form connections around that content. This functionality was particularly meaningful for users in Macha, whose culture was not well represented by mainstream global media. Unfortunately, the end user experience for Facebook users in Macha was quite poor due to performance bottlenecks in the 1 Mbps satellite gateway link to the Internet [107].

This example demonstrates how traditional cloud-based applications and services

may not be amenable to the limitations of edge networks located in challenged contexts. Content generated locally and stored in the cloud must be re-downloaded over a low-performing (and potentially costly) link for other community members to interact with it. Conversely, data collected and content created locally must be uploaded over the same low-performing links to the cloud in order for local community members and organizations to make use of them. Unmodified, this cloud-based model of content collection and sharing prevents communities from truly engaging in their local information economy. In this chapter we discuss existing paradigms in data collection and content creation and how they leave much to be desired in challenged environments. We then present two applications that modify the cloud-based model to empower communities in challenged environments to participate in data collection and content creation.

8.1 Existing Paradigms

Challenges with deployments are often magnified in resource-constrained environments because of insufficient design paradigms.

Uniform Data. Existing routing paradigms often assume that inherent data properties are sufficient to determine the appropriate network technology for data transmission [49, 171]. This assumption overlooks the fact that data is not uniform but instead has two distinctive qualities: inherent qualities *and* contextual qualities. Inherent qualities of data, such as data types and sizes, are independent of the application in use. In contrast, contextual data qualities are necessarily dependent on use scenarios. Examples include data priority, data importance, deadlines, and precedence. Contextual qualities such as an organization's data policy and local laws can also affect how data is stored and transmitted (e.g., private medical records vs public data).

Single-Task Mobile Apps. Resource-constrained environments often lack

enough technical personnel to build and customize information systems. This leads organizations to use productivity software (e.g., MS Excel, MS Word) to create solutions that can be customized by staff having little programming expertise. These tools have been designed for conventional PCs that are poorly suited to these limited infrastructure environments. Although mobile devices are well-suited to scarce connectivity and sporadic grid power, mobile software tools do not yet offer the same range of features as customizable PC productivity tools. Instead, several small apps focused on single tasks are created leading to specialized apps with minimal customizability. Mobile frameworks, such as ODK 2.0 [25], are needed to help organizations customize and refine their apps to their context while maintaining the single-task paradigm. Additionally, with single-task apps there is often limited coordination of system resources making it challenging to conserve resources. For example multiple apps could simultaneously attempt to communicate when connectivity becomes available.

Similar Transmission Cost. Developers often choose a single transport protocol such as TCP/IP or SMS because of systems abstractions and availability of networks for the original deployment location. However, the cost associated with connectivity can vary across different regions creating feasibility issues for deploying applications in varying contexts. For example, a 500MB post-paid mobile broadband subscription in Europe costs 1% of per capita GNI. By contrast, the same subscription costs 38% of the average per capita GNI across Africa [96]. Even as the cost of broadband subscriptions falls globally, an entry-level broadband connection continues to cost over 100% of per capita GNI in less developed countries, as compared to only 1% of per capita GNI in more developed countries [95]. Even in developed regions, there are communities yet to be covered. In the US, broadband coverage on Native American reservations is less than 10% per capita [69], despite coverage of over 70% for the rest of the country. Although technologies with universal connectivity options like satellite uplinks exist, financially

constrained organizations cannot afford them. Restricting data transmission to a single protocol can lead to missed opportunities in optimizing transmission costs based on contextual qualities of data.

8.1.1 Existing Synchronization Tools

Cloud-based data storage and synchronization systems often serve as building blocks for application development. To better understand the performance of existing tools, we evaluated the operation of three popular cloud-based systems: Dropbox [1], OneDrive [2], and Google Drive [3]. We measured performance on Android devices using libraries provided by the cloud services since we are evaluating developer options for creating mobile data applications. Since not all contexts suffer from challenged networks, we chose three cities (Lima, Peru; Kisumu, Kenya; & Lahore, Pakistan) in somewhat resource-constrained countries. Large cities have better infrastructure than their rural counterparts demonstrating best-case scenarios for countries with resource constraints. To provide context for the performance divide between urban and rural areas, we also evaluate the performance differences of an urban city in the U.S. (Seattle, WA) and rural towns in the U.S. (Chowchilla and Pala, CA), which have populations of 600K, 18K, and 1.5K respectively. Service carriers used in the experiments include Claro (Lima), Telenor (Lahore), T-Mobile (Seattle), Verizon (Chowchilla) and AT&T (Pala). File synchronization was measured 15 times per service using 10KB, 500KB, 1MB, 2MB, and 10MB of image files and 1KB, 10KB, 100KB, and 1MB of text files.

Table 8.1 shows network statistics recorded using ping and iperf tools. Cellular networks in Lima and Lahore tend to be highly latent with low bandwidth capacity with Lima experiencing the longest round trip times at 1,173 ms. Seattle was 5 and 30 times

[1] https://www.dropbox.com/home
[2] https://onedrive.live.com/
[3] https://www.google.com/drive/

Table 8.1: Network measurements from various locations

Location	RTT (ms)	Bandwidth (Mbps)	Loss (%)
Lima	1173.0	1.05	0
Lahore	207.0	1.05	0
Chowchilla	154.6	1.69	0
Pala	63.0	6.93	0
Seattle	39.2	11.00	0

faster than Lahore and Lima respectively. Mobile network connectivity in rural US test sites was significantly lower as Chowchilla had 1.69 Mbps (154.6ms RTT) and Pala had 6.93 Mbps (63.02ms RTT) compared to Seattle's 11 Mbps (39.2ms RTT) high speed bandwidth. We emphasize that network availability and performance in Chowchilla and Pala represent a best case for rural connectivity in the US, as measurements were taken from the more densely populated town centers. We also note that poor network infrastructure is not just a problem for developing countries, but for rural parts of developed countries as well.

Results from the performance tests shown in Figure 8.1 reveal that Dropbox performs better than other platforms with all tested file sizes. This is unsurprising as the Dropbox API compresses all files prior to transmission, while Google Drive and OneDrive do not. OneDrive only performed differential synchronization of Microsoft Office files, which excludes many large media files. Google Drive did not use differential synchronization for any file, causing it to use the most bandwidth per file update of the three evaluated options. When accessed via the developer API, Dropbox does not provide differential synchronization, though it does perform data compression prior to transmission. We also note that OneDrive experienced higher variability in file transfer times than Google Drive and Dropbox, with a standard deviation of 5.2 seconds compared to 2.3 seconds for Dropbox and 4.6 seconds for Google Drive. Even though Lahore, Kisumu, and Lima

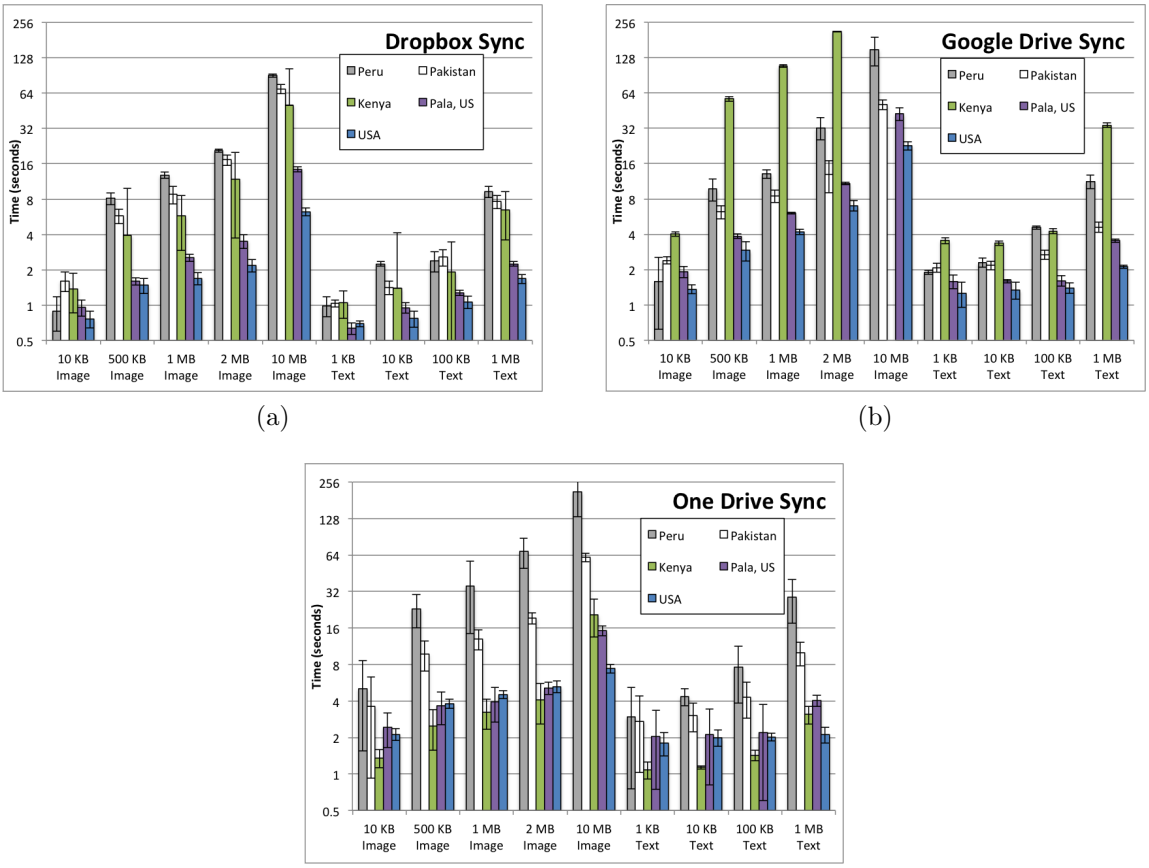


Figure 8.1: (a) Dropbox file, (b) Google Drive, and (c) OneDrive file synchronization performance with varying file sizes using mobile data connection. (Log Scale)

have access to mobile Internet, the performance of the connections were inconsistent and more prone to high latency and limited bandwidth than connectivity in Seattle.

Our measurements demonstrate that common data synchronization platforms experience issues in varying network environments as they all experienced longer file transfer times and greater variability in transfer time in resource-constrained environments. Based on our experiments, Dropbox would be the preferable 'off-the-shelf' solution, followed by Google Drive, then OneDrive. However, there are still issues that these cloud synchronization platforms do not address such as: 1) they lack support to enable organizations to treat data differently based on contextual data qualities; and 2) they are TCP/IP

based and do not allow for alternative connectivity options in challenged network environments.

8.2 Challenges for Information Collection and Creation Mobile Applications

Information collection and creation in challenged network environments poses challenges to the design assumptions of many collection and creation platforms [89]. As mentioned in the previous chapters, telecommunications infrastructure is not ubiquitously accessible for reasons of availability, cost, or capacity. In this section, we draw attention to specific scenarios that demonstrate how data collection and content creation in challenged environments can confront the assumptions of mobile application design.

8.2.1 Example Scenarios

Based on our partnerships and experiences, we outline scenarios that highlight networking challenges and the benefits of leveraging contextual data properties when making data transmission decisions. The examples that we chose were selected to showcase some of the challenges Indian Country faces in participating in activities that assume pervasive broadband access.

Scenario 1: Community Mapping. One of the ways that tribal communities are seeking to practice self-governance and self-determination is through tribal-directed mapping of tribal lands [181]. However, mapping tribal lands can be exceptionally challenging given the lack of ubiquitous telecommunications infrastructure. Data collected might include pictures of the community, GPS coordinates associated with specific resources (e.g., antennas, roads, forests, rivers) and human activities (e.g., travel,

homes, shopping), textual or audio recordings about Indigenous names of places, and signal-strength measurements (in the case of mapping cellular data connectivity). When community surveyors encounter a critical issue with infrastructure (e.g., a mudslide blocking a road or a collapsed bridge), GPS coordinates and relevant information would need to be transmitted as soon as possible to tribal headquarters so the appropriate responders could be dispatched to the area. In the case of routine data collection (which can be very media rich), surveyors' devices will upload collected data to a tribal-owned repository when the cheapest, high-bandwidth telecommunications channel (i.e., WiFi Internet access at the tribal library) becomes available.

Scenario 2: Live News Coverage. Tribal sovereignty is continually challenged by government and corporate interests, as evidenced by a plethora of court cases over the past century [61]. While the rise of social media has helped to bring awareness to these issues [204, 207, 191], tribes have continually emphasized the need for access to infrastructures and platforms that empower them to share information about these issues in real-time [3, 57] so that intertribal and humanitarian efforts can come together to protect tribal sovereignty as a whole. One specific example of this need comes from the *#NoDAPL*[4] movement, wherein the Standing Rock Sioux and fellow water protectors physically stood against the Dakota Access Company and the Army Corps of Engineers in order to protect tribal land and water. One of the significant hindrances for the Standing Rock Sioux to share their perspectives was the lack of mobile broadband infrastructure available at the Oceti Wakowin Camp that functioned as the base of operations for the movement [56]. Without this critical infrastructure, activists had to rely on mainstream media reporting and information shared *post hoc* when water protectors were able to access mobile broadband (reportedly 20 miles outside of Oceti Wakowin).

Scenario 3: Offline-Online Collaborations. In the Information age,

[4] "No Dakota Pipeline".

one big challenge for students living in rural and disconnected communities is the lack of home Internet. Especially at a time when curriculum is increasingly project-based and collaborative [197, 43]. Homework often requires certain Web resources to be accessible from home and students to edit or create Web-based content (alone or in collaboration with each other) [18, 217]. While the move to the reliance on "cloud-based" technologies in education have demonstrated beneficial from a pedagogical perspective, it prevents students without ubiquitious broadband access from participating fully in assignments. As most cloud-based technologies are not designed to seamlessly transition between the connected environment of schools and libraries and the disconnected environments (i.e., students' homes), there is a burden placed on students to proactively download all content necessary to complete assignments offline and to coordinate and merge collaborative assignments in an asynchronous manner that increases the effort necessary to participate in what might seem to be a simple collaborative assignment.

8.3 Submit: A Composable Communications Layer for Mobile Data Collection

Mobile devices often have several built-in transmission capabilities (e.g., GSM, WiFi, peer-to-peer) but lack a flexible framework to systematically adjust to changing network conditions based on an application's deployment requirements instead of simple connectivity available recognition. Data transfer is integral to an application's usage and context making it difficult to create a universal solution to address diverse requirements. This chapter argues for creating a software tool that selects appropriate data for transmission over available network channels and can be customized by deployment architects. Deployment architects are generally non-programmers who adapt an ensemble of off-the-

shelf software to a deployment context. Providing flexible transmission management to deployment architects could improve the feasibility of deploying mobile information systems in challenged network environments by enabling application-level communication optimizations.

Challenges for deploying applications in developing regions have been documented [21] and include: low literacy, limited technical personnel, use of inexpensive multipurpose devices, and context-specific customization. Research initiatives that address some of these challenges have focused on interface design [52, 206] and rapid customizability [88] to produce frameworks such as Open Data Kit (ODK) [25, 88] and CommCare [53]. These frameworks focus on lowering technical barriers to assist organizations in deploying information services in resource-challenged contexts and are designed for disconnected operation. However, ODK leaves decisions about when and how to transmit data to the end-user which leads to possible inefficiencies with respect to transmission costs or deadlines.

In this chapter, we examine options to enable non-developers to adapt their mobile application to various network conditions. To motivate the need for a configurable data transmission tool that operates in sparse connectivity, we discuss challenges with existing paradigms, characterize the performance of popular data transfer apps from different locations, and formalize sources of transmission meta-data into three perspectives. We then propose an (ODK) extension called ODK Submit that uses an organization's deployment parameters to guide communication decisions. Submit enables application-level communication optimization of sparse heterogeneous networks by sending appropriate data over available network infrastructure or peer-to-peer communications. We also investigate and characterize Android peer-to-peer transfer methods to better understand deployment trade-offs for different use cases.

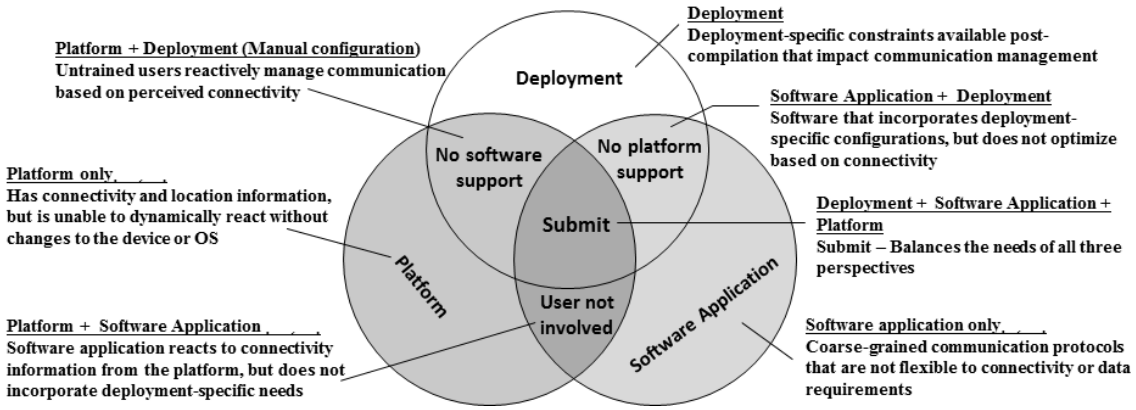


Figure 8.2: Design space of communication solutions for utilization of heterogeneous networks.

8.3.1 Integrating Multiple Perspectives on Data

Data is often transmitted using whatever protocol a software developer selected when developing the software. Dynamic selection of available protocols based on a deployment's context and user location could improve connectivity in challenged network environments. To better facilitate dynamic selection the traditional concept of the TCP/IP Application Layer [180] should be extended to include: 1) metadata from the platform about connectivity; 2) data properties from the software/application developer; and 3) contextual constraints from an deployment architect. The deployment architect is the domain expert who deploys the software in the field and customizes it to meet the needs of their business or organization. The deployment architect and the software developer both provide vital information that is necessary to understand an application's communication context and constraints. Our approach of breaking the application layer into parts is similar to Martins et al.'s approach to coordinating different perspectives on system power[122]. We find their multi-perspective approach to optimizing battery life suitable to optimizing communication resources. As Martins et al. point out: *"the user needs to drive"*; claiming: *"1) The OS cannot always know the resource priorities of all applications; 2) applications cannot always know the functionality priorities of the end-*

user; and 3) users should choose the right level, trading off functionality versus lifetime." Overall we agree with these insights with the exception of focusing on the end-user. Instead, there is often a deployment architect that handles organization-wide restrictions and imposes constraints derived from deployment considerations. The focus on a deployment architect in addition to the end-user is an important distinction, as grouping developer, deployment architects, and end-users into a single concept can make system optimizations difficult. Figure 8.2 shows how Submit aims to combine information from the network perspectives of the platform, software developer, and deployment architect to efficiently manage communication resources to relieve the end-user of communication management.

Platform Perspective. The platform perspective encompasses the device and operating system perspectives on connectivity and mobility. For instance an Android device can: detect the type of available network connectivity, detect device mobility, estimate available data capacity, and estimate the device's geographical location. Location and mobility information enable Submit to possibly infer the duration of a connection and apply regional data policies. However, the platform is unaware of what policies, financial restrictions, and other data priorities a user or organization may want applied. There are numerous works related to a platform-only communication management scheme as multiple communication channels can be dynamically allocated based on availability or bonded to create compound channels with greater throughput capacity[47, 214]. While Submit dynamically schedules traffic based on channel availability, it does so without modifying the underlying platform to combine or bond channels.

Application Developer Perspective. The application developer perspective encompasses issues relating to the functionality of a mobile application. A developer understands the inherent properties such as the type and typical size of data that can help Submit develop an appropriate cost model. Unfortunately, a developer may be biased

towards a particular communication medium and may not bother including functionality to support alternatives such as: local off-line storage to support disconnected operation or transmission of summary data over SMS. Thus, developers constrain communication resources via software design and protocol selection. Some examples of developer limitations include apps that communicate over 2G and 2.5G networks exclusively rather than 3G data networks[165, 210]. Likewise, a single protocol limits what an app can effectively communicate, for example, transmitting binary over SMS is non-optimal. Furthermore, a developer likely does not fully understand how a future user may want to deploy the application in varying context with limiting data policies and budgets.

Application Deployment Perspective. The deployment perspective encompasses issues relating to contextual deployment requirements that should be incorporated by a deployment architect, as the dynamic contextual information is not available when the developer compiles the software. A deployment's requirements can provide important meta-data including information prioritization, and financial restrictions that may change during the lifetime of the project. ODK tools are designed to be general-purpose and have been deployed in a variety of settings including public health, environmental conservation, and census applications. These domains have different needs and real-time information may change how data is transmitted. For example, in a health application, a CHW may find a patient needing immediate referral to a care facility. This message is urgent and should be sent with a different priority than updating a healthy patient's medical record. Usage context is not predictable by the developer nor can the platform impose that one particular channel be used as that channel may not be available. Depending on the urgency of the data it may be necessary to send the same data over multiple channels to ensure delivery or possibly reach multiple destinations. Submit seeks to remove the user burden of actively monitoring the status communication events.

Overlapping Perspectives. There are points where each of these perspectives

overlap and interact. The most commonly combined perspectives are those of the platform and software application. Opportunistic off-loading approaches combine the connectivity awareness of the platform with software application protocol decisions [12, 86, 118]. While this approach provides more guidance than either platform or software developer perspectives alone, it results in coarse-grained communication automation. In contrast, web applications exemplify the absence of the platform perspective [134, 140]. While there is value to platform-independent systems, platform information about connectivity and location is necessary for maintaining historical connectivity models. Submit's design incorporates information from the different perspectives and attempts to hone how and when information is transmitted in challenged networking environments.

8.3.2 Related Work

Previous research has explored leveraging a variety of networks for data transmission [11] and splitting data over multiple networks based on cost and availability [12, 86, 118]. While work exploring simultaneous data transmission over multiple interfaces has been shown to improve mobility, power efficiency, and network capacity [11], our work focuses on selecting a single network from a heterogeneous combination of different transmission opportunities in a manner specialized to the deployment context. Submit is most similar to work that focuses on identifying the best type of network for data transmission given various contexts and policies. MultiNets [138] proposes real-time switching between different network interfaces on mobile phones using policies based on power, data offloading, throughput, and latency. However, policies are configured by the user and are applied to every app that uses the device. In contrast, Submit provides a library that allows deployment architect to configure policies that will only be applied to apps relevant to the application. In this way, Submit is most similar to Delphi [49], a

transport layer module that selects the most appropriate network for data transmission given policies set by applications. However, Delphi assumes operation in an environment with ubiquitous connectivity and focuses on the transport layer not the application layer. While it attempts to provide a systematic evaluation of various networks for data transmission, it does not address many of the issues of developing contexts including intermittent connectivity and regional pricing policies. Also in contrast to Delphi, Submit also uses information about data (e.g., time-sensitivity, importance) to identify the best method for transfer. Haggle [171] is another solution that separates application logic from inflexible pre-programmed transport bindings so that applications can communicative in dynamic networking environments. Haggle and Submit both provide API's to developers but Submit goes further and provides constructs for deployment architects who are not programmers to adjust their composable mobile information system. Another issue with Haggle is that it proposes a general form of a naming notation to allow for late-binding that is independent of the lower-level address. While a good idea, the infrastructure is not currently available to support some of these assumptions. In contrast, Submit is designed to work with existing infrastructure to enable applications to "just work" in different environments with limited infrastructure. Our work distinguishes itself from previous research in that it seeks to provide a network management module at the application layer that enables flexible control to an organization deploying a customizable app in areas of limited connectivity.

8.3.3 Submit System Architecture

Submit is an Android service that coordinates data communication by providing channel monitoring and transmission scheduling mechanisms to Android apps. For the purposes of this section, the term client app refers to any Android app that binds to

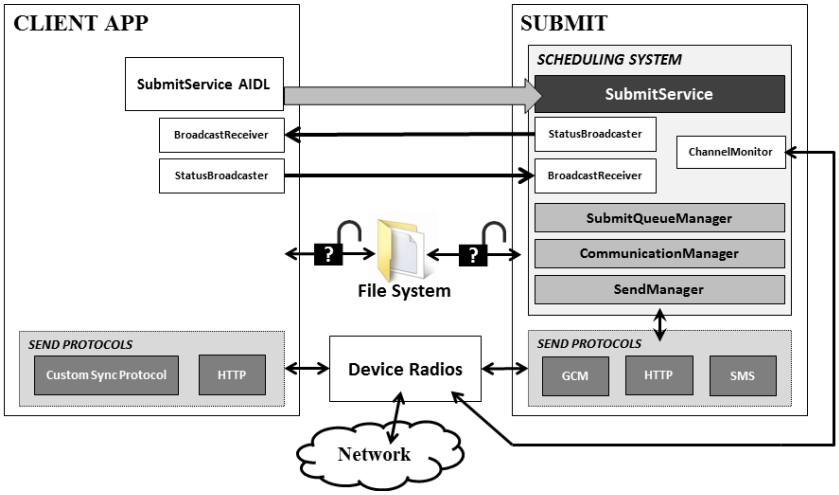


Figure 8.3: Architecture diagram showing how Submit interacts with a client app and Android system resources

Submit's Android service. Submit provides software developers with an interface that abstracts communication channels and flexibly handles data ownership and application-specific synchronization issues. Submit is designed to separate application logic from the network routing logic with a communication system that provides extensibility in terms of: 1) adding transmission channels; 2) modifying transmission channel selection; and 3) handling complex data ownership and application-specific synchronization issues. Submit's service API exposes communication scheduling mechanisms that client apps use to either 1) delegate responsibility of transmitting data to Submit; or 2) register to receive notifications when appropriate network channels are available. When using Submit for notification purposes, a client app takes responsibility to transmit its data with its own possibly proprietary or complex protocol.

Client apps specify two types of objects to interface with Submit: 1) the DataPropertiesObject and 2) the SendObject. The DataPropertiesObject contains metadata that describe the data to be transmitted. The properties supplied include: data size, data urgency, data fragmentability, and reliability requirements. The inherent properties are derived from the perspective of the developer representing "normal" sized data for the

client app as an app that primarily transmits responses to survey questions has larger "normal" data than an app that primarily transmits simple reminders the size of SMS messages. By obtaining the software apps perspective on data size, Submit is able to calibrate its routing mechanisms to best handle the common communication case on a per app basis and select appropriate channels. The SendObject contains a list of DestinationAddresses that define the type of transport as well as the necessary parameters to use for the transport. For example, to utilize HTTP `POST`, the DestinationAddress would contain a URL; to utilize an SMS channel the DestinationAddress would include a phone number. By implementing DestinationAddress as an abstract type, Submit is extensible to various communication protocols. SendObjects also contain the file path to, or string representation of, the data to be sent on behalf of the client app.

Shared data between Submit and client apps could create race conditions as apps are often dependent on their internal data stores being correct and consistent. Submit addresses ownership issues by removing ambiguity through the assumption that the client app owns the data until it explicitly grants Submit temporary ownership rights when it delegates transmission responsibility to Submit. If an app only provides a DataPropertiesObject which has no pointers to the data, Submit assumes the app is maintaining ownership of the data as the client app is only asking for a notification of when to send the data. In contrast, if the client app provides a SendObject containing the actual data or a pointer to an accessible external file, Submit retains ownership of delegated external data until it notifies the client app with the final status of the transmission. Since both the client app and Submit can be responsible for sending data, they must communicate the success or failure of data transmission. Broadcast intents are used to synchronize the sending status. When Submit is responsible for transmitting the data, it broadcasts the status of the communication exchange to the client app. Likewise, if a client app has been notified it is the appropriate time to send the scheduled data, the client app

broadcasts the transmission result status to update Submit's internal state.

If the client app delegates responsibility for sending to Submit, the SendManager selects an appropriate network to transmit the data based on the DestinationAddresses and the protocols suited to the available network. By providing multiple client libraries that implement various protocols, Submit increases a client app's ability to communicate using various protocols without requiring expansion of the client app's code base. Currently implemented communication protocols include HTTP/SSL and SMS.

Submit's CommunicationManager is responsible for determining whether an available channel is appropriate for submitted data. The CommunicationManager gauges an available channel's bandwidth capacity and costs. The ChannelMonitor listens to system broadcasts for changes in connectivity, including WiFi events, ad hoc communication opportunities, and cellular events. It reports back the current state of connectivity to the SubmitService when a change is detected. The SubmitQueueManager iterates continually over the pending data that needs to be transmitted (described by DataPropertyObjects). With each pass through the queue, it updates the state of each pending data item based on the results from Submit's protocol modules.

8.3.4 Experiments

Splitting Data Transmission. To measure the baseline impact Submit has on a client app's communication performance, we integrated a simple file upload app with Submit to evaluate network usage and latency. The test involved the client app using HTTP POST to upload data from a client to a server with and without Submit. Performance was measured in two scenarios: WiFi only and 3G only. The tests were performed on a Samsung Galaxy running Android 4.3 using either 3G or WiFi networks. The results in Table 8.2 show the average latency for each file size for the ten uploads. As

Table 8.2: Average latency for a client app sending data using Submit and without using Submit.

	WiFi w/ Submit	WiFi w/o Submit	3G w/ Submit	3G w/o Submit
10 KB	0.12s	0.10s	0.59s	0.47s
100 KB	0.40s	0.28s	1.66s	1.28s
1 MB	2.53s	2.03s	10.93s	7.86s
10 MB	22.55s	20.58s	83.95s	80.35s

expected, there is slight latency overhead associated with Submit due its use of remote procedure calls and broadcast intents to communicate with a client app for each uploaded file.

Submit's latency additions are counterbalanced with its ability to minimize network usage according to user preferences. To verify the reduction of cost for network usage a small experiment was performed for ten minutes where WiFi was disabled, leaving only 3G accessible. In this experiment, the client app uploaded a randomly selected file between 5 KB and 100 KB from a directory of files. For the client using Submit, a threshold value was set that prevented any file over 7 KB from being sent over a mobile broadband network. After ten minutes, the WiFi was re-enabled for 10 minutes. After the entire 20 minutes the number of packets sent over WiFi was compared to the more costly 3G network. The client without Submit sent over 380.6 KB over 3G whereas the client using Submit only sent 12.8 KB over 3G. Thus, Submit selectively uses one network while waiting for cheaper channel to become available.

To understand Submit's effects on deployment scenarios a small sampling of 85 actual site visit records were used to calculate data transmission reductions for the Site Visit scenario described in section 8.2.1. Table 8.3 shows the calculated average reduction if Submit managed ODK Collect's data submission process. By simply separating

Table 8.3: Reduction of transmission if record is split by data type or data priority for site visits scenario

	Bytes	Percent
Avg. total record transmission size	330,773	100.00
Avg. data size	1,213	0.37
Avg. photo size	329,217	99.53
Avg. priority data size	343	0.10

transmission of the text portion of the data and delaying the transmission of the photo's binary data until the device is in free WiFi range would mean 0.37% of the total data would be transmitted via cellular and 99.53% of the data would be transmitted over WiFi. Using Submit's concept of data priority could further reduce cellar transmission to 0.1% of total bytes by only transmitting the information about medication inventory.

Peer-To-Peer Transmission. We evaluate the performance of peer-to-peer transmission methods by comparing 5 methods of transferring data between Nexus 7 devices positioned 0.5 meters apart running Android 4.4.4. WiFi Direct and Bluetooth were tested using transfer sizes of 1 KB, 10 KB, 100 KB, 1 MB, 10 MB, 100 MB, and 1 GB of data. NFC transfer sizes were limited to data transfer times that were feasible for generic peer-to peer use (large transfers took too long). NFC with Bluetooth was tested up to 10 MB, while NFC only was tested up to 100 KB. Additionally, peer-to-peer transfer using QR codes was tested by having one Nexus 7 display QR Codes on its screen while another Nexus 7 read the QR codes with its built-in camera from 0.3 meters away. The ZXing library was used to generate and scan the QR Codes. While QR Codes specifications state transmission up to 4 KB of data [50], our results show that transfers are unreliable past 1KB of data. The time it takes to scan a QR Code is fairly consistent as the size of data increases, but the error rate increased to over 60%

for file sizes larger than 1 KB.

Bluetooth and NFC were the fastest transfer options for smaller amounts of data as shown in Figure 8.4. As the data size increases WiFi Direct emerges as the fastest mode of transfer. Until data sizes exceed 100 MB, the total time for WiFi Direct remains essentially constant because establishing the connection dominates the transfer time [29] as shown in Figure 8.5. WiFi Direct is a realistic choice for data on the order of 1 MB or larger, for anything lower than 1 MB Bluetooth may be a better option. Figure 8.4 also shows that NFC only is significantly slower than Bluetooth enabled NFC[5]. QR scanning had the largest variance in the duration of file transfer. While increasing the error correction level of the code can help remedy this issue, for data sizes close to 1 KB the QR Code was too dense to accurately and consistently be scanned.

Since battery life is important in disconnected environments, we evaluated the power performance of WiFi Direct, Bluetooth, and QR codes. For each test a connection was established between two fully charged Nexus 7's and data was continuously transmitted from one device to the other until the sending device's battery was depleted. For consistency the device screens remained on at all times since the QR method requires the screen to be on and active usage of devices would cause the screen to be active some percentage of the time. The experiments revealed that despite being able to leave the device in airplane mode, the QR Code scanner consumed more battery than traditional data transfer methods. The QR scanner took 6.8 hours to drain the battery to a 10% level while WiFi direct transfer only lasted 0.33 hours longer. In comparison it took about 9.3 hours of Bluetooth transmission to drain the battery to a 10% level. The results suggest that the power required to continually use the camera and process bar codes resulted in greater battery consumption over time than WiFi or Bluetooth transmission.

The main factors for selecting a peer-to-peer method include transfer time and battery

[5]http://developer.android.com/training/beam-files/

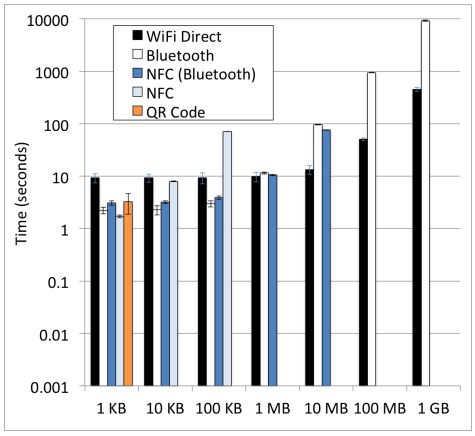


Figure 8.4: Data transfer times associated with peer-to-peer technologies with different file sizes. (Log Scale)

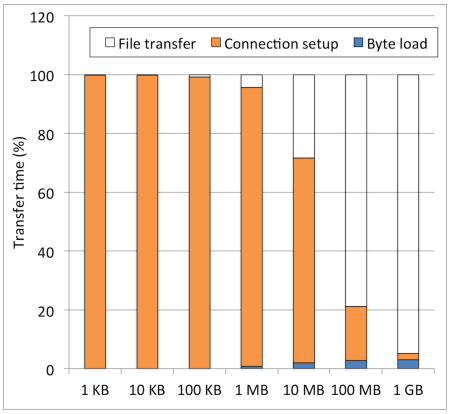


Figure 8.5: Percent of time spent in different phases of WiFi Direct transfer. Connection setup time dominates small file size transfer.

efficiency. Per byte, WiFi is more battery efficient, but within the range of 100 KB to 1 MB, Bluetooth is faster. In the case of forest inventory workers, opportunity to charge the devices is the limiting factor and WiFi should probably be selected, since it is more efficient per byte. For the CHWs, avoiding the slightly more cumbersome connection process of WiFi might be more important. The main disadvantage of both Bluetooth and WiFi Direct is the difficulty for the user to confirm which device they connected to. While this may be less of an issue in a forest inventory setting, it is one of the primary concerns in a clinical setting. With both Bluetooth and WiFi Direct, someone

attempting to steal data can spoof their device name and MAC address, potentially deceiving a user. NFC and QR Code communication allows users to visually clarify that the correct device is receiving the data. This can be an important advantage when the information is confidential such as medical data. The results show that the QR Code scanner is slower and less power efficient than NFC with Bluetooth. However, there is the possibility of hand-to-hand contact from using NFC when the two devices are brought close together to establish the connection. Hand contact could be a disadvantage in a remote clinical setting where hygiene practices might restrict such contact. A key advantage of Bluetooth over WiFi Direct is the ability to pair devices ahead of time, allowing users to more confidently send their data to the correct person. However, if NFC is not an acceptable option due to hand-to-hand contact or data size, white-listing Bluetooth devices could increase security in a clinical setting.

Usability of Peer-To-Peer Transfer. To understand the overhead of using different peer-to-peer modalities we conducted basic usability tests with 22 participants. The participants' ages ranged from 18 to 56, with a mean age of 25. After initial demographic information was collected, participants were given a short training session on how to use Submit's manual peer-to-peer transfer screen. Participants were then given a list of ten 1KB transfers tasks to complete, one sending and one receiving for each of the five transfer methods. The order of the task list was randomized across participants so that each transfer method appeared with similar frequency at each position. The first two tasks for each user used a specific transfer mode (e.g., send using NFC, then receive using NFC). After the first two tasks, participants were asked to complete the NASA TLX[6] form to rate the difficulty of the specific transfer mode. Once completed, participants proceeded with the eight remaining tasks. After the ten tasks were completed, a semi-structured interview was used to solicit feedback about the most confusing part of

[6]http://humansystems.arc.nasa.gov/groups/tlx/

the transfer process and to help identify possible improvements to Submit's peer-to-peer transfer. Participants were also asked to rank the five transfer methods based on ease of use (on a rank scale from 1 to 5, where 1 was the easiest transfer method to use and 5 was the most difficult), and rank them based on efficiency (on a rank scale from 1 to 5, where 1 was quickest transfer method and 5 was the slowest method).

Usability results confirmed the results from transmission performance. Users found that using a QR Code to transfer data was both the least efficient ($p < 0.001$) and the most difficult method ($p < 0.001$) to use[7] with a mean efficiency rank of 4.7 and a mean difficulty rank of 4.6. In three cases, users were unable to successfully scan the QR Code due to the lighting conditions of the room. Users also found that Bluetooth and WiFi Direct were the fastest ($p < 0.001$) and easiest ($p < 0.001$) to use. WiFi Direct had a mean efficiency rank of 2.3 and a mean difficulty rank of 2.3. Bluetooth had a mean efficiency rank of 1.9 and a mean difficulty rank of 1.8.This slightly differs from the data gathered during channel testing. Due to the longer connection setup time, WiFi Direct should have been outperformed by all four of the other transfer methods. This discrepancy can be explained by the fact that users had to select more information during the user testing. The time it takes to select all this information causes the actual speed of transfer to matter less when compared to the overall time spent using the application. Additionally, the time it takes to move the devices together for NFC, and the time it takes to line up the devices for QR Code scanning was not accounted for in the evaluation of performance. These usability issues dramatically increase the overall time it takes to transfer using NFC and QR Codes. Testing results showed QR Codes take an unreliable amount of time due to user and environment conditions such as reflections due to lighting. However, based on the data collected from the NASA TLX form, there was no significant difference between the average level or type of stress users experienced while using the

[7]Significance calculated using the Kolmogorov-Smirnov Test.

different transfer modes ($p = 0.57$)[8]. This implies that while users do prefer some transfer methods above others, the difference between them is relatively small compared to the overall ease.

While users found Bluetooth and WiFi Direct to be the fastest and easiest methods and QR Codes to be the slowest and most cumbersome, results were more varied for the NFC options. Some users liked the strong visual and physical cues associated with holding the devices together and touching the screen to beam the data between devices. Users also appreciated that the receiving NFC user does not have to select any options since the parameters, such as from whom you are receiving, are all automatically inferred when you hold the devices together. Other users felt uncomfortable with the inevitability of hand to hand contact that comes from holding the devices together with NFC. Other users expressed concern surrounding potentially dropping the device. Most users held the tablets together with one hand, and tapped the screen with their other hand. They thought NFC was slightly more inconvenient than other mediums, but would not mind using it.

8.4 Empowering Localized Content in Rural Schools

Research efforts in the Information and Communication Technologies for Development (ICT4D) community focus on understanding, extending, and improving connectivty in developing regions. These efforts are particulary relevant for the rural contexts of Indian Country where sparse population densities make it difficult to subsidize broadband infrastructure. Currently, centerpieces of this work involve specialized Web caching [205, 38] and extending last mile connectivity via new network protocols [64, 150]. These solutions seek to improve information accessibility for rural communities, with the ex-

[8]Significance calculated using the Kruskal-Wallis Test.

pectation that information access will improve health, education, and economic sectors of these communities. While these solutions make steps towards improving access to external information sources, they do not lend themselves to fully satisfying the nuances of information access in rural educational contexts. This is because solutions focused on information access fail to privilege and integrate Indigenous knowledge as part of a greater collection of global knowledge. Consequently, information systems for education become based in Westernized concepts of "exclusive expert systems" and the "monoculture of best practice" rather than more effective "holistic knowledge systems" that allow for a diversity of perspectives [194, 22].

In this chapter, we present VillageShare, an architecture designed to privilege the sharing of local content generated in rural schools. Previous work reveals a high locality of interest in media access patterns in rural contexts. VillageShare exploits this locality of interest not only to improve the accessibility of relevant information, but to improve the sharing of content that is of likely interest to the surrounding community. VillageShare counterracts limited upload capacity in rural communities by serving locally generated content from local content servers before sharing global copies with external services. This maintains high accessibility for local users, and eventual accessibility from global contexts. By keeping locally generated content highly available, we imagine that the incorporation of ICTs for rural education in Indian Country will be more relevant to students utilizing these technologies. Additionally, VillageShare seeks to empower information sovereignty by providing an explicit sharing model. This way, content must be explicity shared via a social network overlay, maintaining a user-controlled differentiation between public content and private content.

8.4.1 Related Work

Distributed Storage Systems. There are numerous systems that provide distributed storage services for various network environments. We focus our discussion on systems that either focus on reducing bandwidth requirements or that emphasize availability.

Modifications to NFS have reduced bandwidth requirements, but continue to prefer strong consistency over high availability in the face of system partitions [133]. This is undesirable for storage systems that exist over a network that is subject to regular link or power outages. Coda [111] extends AFS to support partitions by providing offline availability. In contrast to VillageShare, Coda is a centralized system, wherein a central server maintains a first class replica of content and clients maintain only second-class replicas with inferior persistence, distribution, completeness, and accuracy.

Amazon's Dynamo [46] uses a key-value store to maintain availability in the face of network partitions. Most relevant to VillageShare is that Dynamo seeks to provide an "always-on" experience for users, sometimes at the sacrifice of consistency in certain failure scenarios. However, Dynamo is designed for the data center environment whereas we focus on resource-limited contexts with wider areas of distribution.

For mobile environments, the Bayou architecture [182] maintains strong consistency over varying levels of connectivity by providing mechanisms for full update roll back and reapplication in the presence of conflicts. This approach assumes the possibility for roll back in all scenarios, which is undesirable where human actions may need to be undone for the sake of consistency. Likewise, Bayou assumes that all data is replicated at every node where VillageShare takes a more localized approach to data storage.

Ficus [85] provides an optimistically concurrent peer-to-peer file system extension to NFS. The main disadvantage with Ficus is that it requires update conflicts to be resolved

before a file becomes available, potentially reducing availability if a conflict cannot be resolved immediately. VillageShare does not focus conflict handling to a per-update level, nor does it render files unavailable in the presence of unresolved conflict, thus preferring local availability to global consistency.

Systems that maintain particular focus on distributed storage in developing contexts include TierStore [48], COCO [173], and Kwaabana [108]. TierStore is a distributed file system that focuses specifically on distribution over wide-area, intermittent, bandwidth-limited networks. It uses the Delay Tolerant and Networking overlay network and the publish/subscribe-based multicast replication protocol. COCO allows for offline operation by queuing upload and download requests and synchronizing them with the central shared repository upon reconnection. Most similar to our work is Kwaabana, which forms the basis for the work described in this chapter. Kwaabana is a file sharing system that facilitates reliable content sharing among rural users as well as between rural users and external users on the Internet. Ultimately, Kwaabana seeks to reduce traffic impact on the Internet gateway link to help minimize the impact large uploads and downloads have on real-time traffic. Our work maintains the goals of Kwaabana, but extends it to enable content sharing between multiple rural communities.

Content Access and Generation in Developing Contexts. Delay Tolerant Networking (DTN) provides a store-and-forward protocol for moving content through a network that does not guarantee end-to-end connectivity. DTN provides a method of connection where there is no traditional data connection–often by opportunistic courriers. Existing systems that utilize DTN include the Wizzy Digital Courrier [160] and DakNet [153], among others [172, 7]. While these solutions do provide a means of access in networks with intermittent connectivity, they are highly asynchronous and focus more on global content access than localized content access.

Several systems are designed to function completely offline, so that localized content

generation is made possible despite a lack of connectivity. Amongst these solutions are Digital Green [73] and Digital Doorway [83]. These solutions are particularly relevant to VillageShare for their educational emphasis. Digital Green is built on COCO and seeks to create a repository of locally produced videos focused on sharing information for improving health, farming, and livelihood. While VillageShare shares the goals of these systems for creating highly available content generation and sharing opportunities, it views sharing as an event between individuals rather than between an individual and the entire community.

8.4.2 VillageShare Architecture

The VillageShare system consists of a logical core, a social networking layer, and a multi-server support layer. VillageShare servers act as local content repositories that are fully functional during times of disconnection. Upon connection, VillageShare servers located close to a high-bandwidth Internet link act as coordinators between VillageShare servers placed in less-resourced environments.

Core. The VillageShare core builds on ownCloud [147], a "self-controlled free and open-sourced" cloud solution. We use ownCloud for several reasons. Most importantly, it is an open source technology, which makes it easily customizable and localizable. It provides simple user interfaces for file management as well as system administration. Additionally, it provides APIs to facilitate app integration for further customization.

The ownCloud Web server uses the model-view-controller (MVC) paradigm: template, server logic, and data.

Server Logic. The main VillageShare server logic is implemented in PHP. Primary objects within the server logic are users and files. The bulk of the server logic is inherited from ownCloud v. 6; in addition to file management, ownCloud v. 6 provides support

for localized data backup, user and system management via a graphical user interface, and local conflict handling.

Interfaces. For the Web application, user interfaces are implemented using AJAX. The three main user templates provide interfaces for user login, file management, and social network management. In addition, administration templates provide interfaces for application management, user management, and system management.

There are three main APIs utilized by the VillageShare core: the OC API, the OCP API, and a REST API for Android clients. For use with template components, ownCloud provides the OC private API. This includes callback functions for creating and managing user and file objects. The OCP public API provides callback methods for third-party custom applications. This API provides a subset of the functionality provided by the OC API. In addition to the APIs provided by ownCloud, we also implement a limited REST API to support Android mobile clients, including functions for sharing files, requesting friendships, and registering users.

Storage. For each content server, VillageShare uses a MySQL database to manage the relationships between users and files. In Figure 8.6, we list the most relevant tables to VillageShare. In addition to the database, VillageShare maintains a file directory for each user it hosts. Again, Figure 8.6 illustrates this directory structure.

Social Networking Layer. Explicit sharing is a critical design factor for VillageShare. This is different from sharing via a collective commons where all users of the system have access to all content hosted at a central repository. ownCloud minimizes the scope of sharing using an open sharing model where any user can share files with any other user hosted on the server. This model of sharing empowers file owners to explicitly select the scope of access to their content. However, this sharing model is inadequate for a school environment where all users may not want to be accessible as recipients of files from users they have no relationship with. Without giving users control

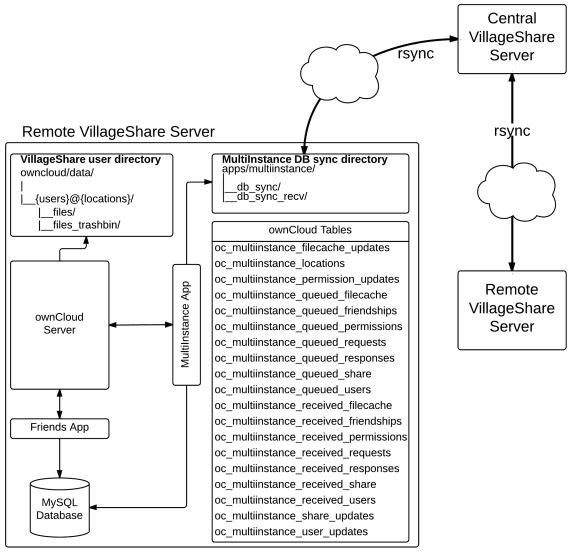


Figure 8.6: VillageShare architecture.

over their receiving scope as well as their sharing scope, users are vulnerable to receiving spam content. Therefore, explicit sharing is sharing between parties where the sharer explicitly decides to share specific content with a specific scope of users and the sharee explicity agrees to receive content from some subset of users. We ensure explicit sharing by imposing the notion of a social network into VillageShare. This is implemented as a third-party ownCloud app that provides an interface for users to request, accept, and delete friendships. Once a friendship has been established, the users comprising the friendship are able to share files with one another.

Multi-server Support Layer. While both the Core and Social networking layer exist at the local level, we extend their scope with the Multi-Instance app for ownCloud. The Multi-Instance app allows multiple VillageShare instances to share files in a bandwidth-conservative manner. We use a centralized control topology, wherein rural VillageShare instances synchronize via the control of a central server. This topology is shown in Figure 8.6.

Transportation. VillageShare uses `rsync`, a network protocol for incremental re-

mote file synchronization. We find `rsync` to be advantageous for several reasons. The protocol minimizes bandwidth requirements by using delta encoding to determine which contents must be sent over the network for synchronization. Another critical feature is `rsync`'s robustness to network outage. If the network link fails between two servers in the middle of a file transfer, `rsync` will resume transfer from its last point of success, reducing bandwidth overhead for content retransmission. Additionally, `rsync` allows for encrypted transfer via SSH[9].

VillageShare prevents aggravating congestion using time-shifting techniques. If large files are scheduled to traverse the gateway link during peak traffic hours, they are delayed from being sent over the network until off-hours when they will make less impact on other traffic.

Synchronization. A subset of the VillageShare database tables are synchronized between rural VillageShare instances and a central controller instance. We experience a tradeoff between robustness to equipment failure and minimizing bandwidth requirements. Equipment failure is a documented challenge of deploying technical systems in rural developing contexts [21]. Likewise, limited bandwidth capacity is another significant challenge for these systems. Thus, creating full backups of all content and all database tables hosted by a rural VillageShare instance would be inconceivable due to the bandwidth requirements. We address this tradeoff by providing a backup of database tables required for content maintenance and sharing. Even though ownCloud provides database tables for maintaining app configuration, file versioning, deleted files, and system preferences, only user profiles, friendships, file metadata, and file contents are backed up at the central VillageShare instance. This is done by storing data for remote backup in queued database tables. When it is time for synchronization, files are synchronized between remote synchronization directories. Remote servers only maintain a single directory that

[9]Secure Shell protocol.

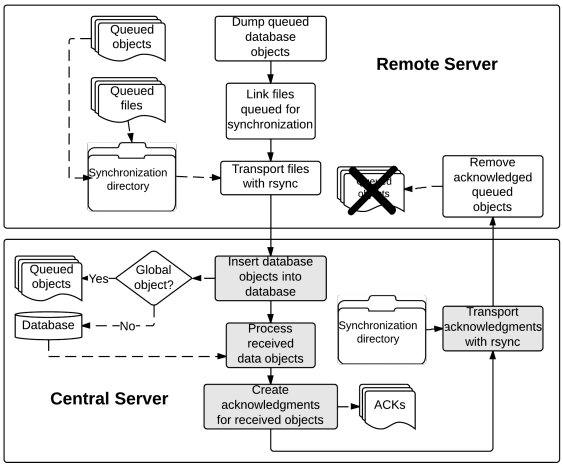


Figure 8.7: VillageShare synchronization process.

corresponds with the central server they coordinate through. Central servers maintain synchronization directories for each remote server they coordinate.

We also note that synchronization occurs in a time-driven manner. The `rsync` utility is called as a `cron` job every N minutes. We discuss potential alternatives to the time-driven model in Section 6.3.

We illustrate synchronization in Figure 8.7. The process of synchronization operates as follows: **Step 1** *Remote server*: Queued database tables are dumped as MySQL `INSERT` statements into a file in a designated synchronization directory. **Step 2** *Remote server*: Files recorded in the queued filecache database are symbolically linked to the synchronization directory. **Step 3** *Remote server*: A timestamp is generated and recorded to the lastUpdated.txt file in the synchronization directory. **Step 4** *Remote server*: `rsync` is used to transport files all files in the synchronization directory across the network to a receiving synchronization directory at the central VillageShare instance. **Step 5** *Central server*: Insert all MySQL files from the synchronization directories into received database tables. **Step 6** *Central server*: Process each received database object. If the received object has no correlated object already in the central server's database,

a new record is created. If a correlated object already exists, it is updated with information from the received object. If received objects correspond to friendship requests or file shares destined for another remote server, the central server creates corresponding queued objects (as described in Step 1) that it places in its synchronization directory. **Step 7** *Central server*: After each received object has been processed and entered into the central database, an acknowledgement file is generated with the same timestamp corresponding to that generated in Step 3. The acknowledgement file contains MySQL `DELETE` statements corresponding the the queued objects that the remote server transferred to the central server in Step 4. **Step 8** *Central server*: `rsync` is used to transport all files in the synchronization directory across the network to a receiving synchronization directory at the remote VillageShare instances under its domain. **Step 9** *Remote server*: If the timestamp on the acknowledgement file corresponds to the last recorded timestamp in lastUpdated.txt, the MySQL `DELETE` statements are executed, effectively removing all queued objects that have been acknowledged and processed by the central server.

When there are no queued objects at the remote server or the central server, the two servers are considered to be synchronized.

8.4.3 Evaluation

System resilience to failure is critical given our focus on rural developing contexts. Thus, our primary goal in evaluating VillageShare is to assess its robustness to network failures, power failures, and poor link quality. In order to evaluate the system, we deploy a three node test bed in the lab using the Linux `Traffic Control` (tc) utility to simulate long distance wireless links between nodes. We vary packet loss rates and network latency throughout the evaluation, but we maintain a bandwidth of 1.54 Mbps between nodes as

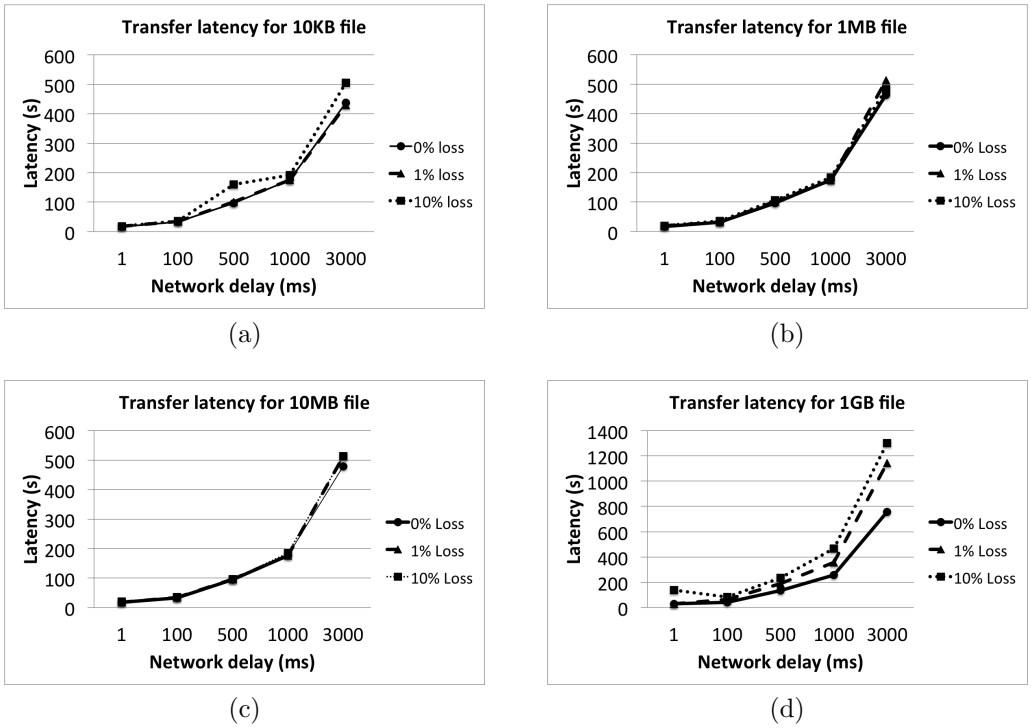


Figure 8.8: Upload latency for (a) 10 KB, (b) 1 MB, (c) 10 MB, and (d) 1 GB text file.

this best simulates a standard T1 link.

Fault Tolerance. We leverage the `rsync` utility to provide fault tolerance in the face of network partitions caused by link and power failures. If a remote VillageShare server is disconnected from the Internet, it will continue to function as a local content server until it reconnects, at which point it transfers all data it queued during the period of disconnection according to the limits imposed by our time-shifting mechanism. If power failure renders a server unreachable, it is unavailable as a local content server. During periods of local outage, users can access files backed up to the central server, although this access method imposes large latency penalties if user's wish to download any of their content hosted at the central server. As soon as power is restored to the local content server, queued data from other VillageShare servers are transferred to the restored server and the system-wide synchronization is restored.

Latency. Latency of file upload is assessed with respect to file size, networking latency, and loss rate. We upload files of varying sizes, from 10 KB to 1 GB, under various conditions. Our results can be seen in Figure 8.8, and they reflect the mean measurements recorded for five uploads at each set of network conditions.

As seen in Figure 8.8c, for file transfers of 10 MB and smaller, varying packet losses up to 10% loss does not significantly impact transfer time. This is due to `rsync`'s ability to resume a transfer from it's last point of success, rather than restarting an entire transfer entire after a certain timeout period or number of packet retransmissions. We note that for 10 KB, 1 MB, and 10 MB sized files, the impact of loss rate is not as significant as it appears to be for larger 1 GB files. This confirms our time-shifting technique for delaying larger file transfers until after peak hours where they may suffer from increased loss rates due to higher levels of congestion. Additionally, we note that for each set of network conditions we measure, upload latency for 10 KB, 1 MB, and 10 MB are all within ten seconds of each other. We notice that for 1 GB files, transfer times can be up to two times longer than for smaller file sizes under very lossy conditions. Again, this validates our time-shifted approach to large file uploads and helps gives us insight into reasoning about an appropriate threshold to use for differentiating large and small files.

8.5 Discussion and Conclusion

8.5.1 "Disconnected" as a Common Case

Most content creation and data collection platforms assume Internet connectivity. For instance, many social media platforms (a major subset of Web content creation platforms) will not function when there is not network connectivity (even if content has been previously downloaded or if a user wants to compose a content to post when connectivity

is re-established). Our usage scenarios and demonstrate how these design assumptions can clash with the realities of users who are collecting and creating information in areas where being disconnected from the Internet is more typical than being connected. While our work seeks to provide something of a buffer between users and these design conflicts, existing content creation and data collection platforms might do well to consider usage in disconnected or poorly connected areas as particular modes of operation that require special handling.

8.5.2 Easing the Burden of Designing for Heterogeneous Networks

Configurable mobile data transmission frameworks can enable deployment architects with limited programming skills to adapt mobile devices to meet diverse application requirements in challenged network environments. High-level data and networking abstractions can improve mobile app design paradigms by making single-purpose apps more malleable to resource-constrained contexts where issues such as affordability, infrastructural constraints, institutional capacity, and technical support are nontrivial. Our work with Submit improves deployments by supporting variations in deployment contexts in a systematic manner. Our work demonstrates how abstractions that decouple data and connectivity can enable application-level optimization of sparse heterogeneous networks. However, there is a need for work that investigates the extent to which these abstractions should be made transparent to users and app designers. For instance, future research might ask: Would it be helpful for a designer to be able to account for "offline" scenarios in their app and change application flow accordingly? What benefits are there for a user to know if they are offline or online and how might this change their behavior and perception of their experience with an app?

Experiences from the field highlight the fact that the selection of appropriate technologies for data transmission involves accounting for inherent data properties, contextual data properties, and net- work properties. Submit enables deployment architects to shape the communication priorities of the components comprising a larger application. Several deployment scenarios of interest benefit from peer-to-peer connectivity when centralized infrastructure is unavailable for data transmission. Which peer-to-peer transfer method to use should also be based on inherent properties (e.g., data size), contextual properties (e.g., data importance, security), and battery constraints. Our experiments showed the most significant barrier to peer-to-peer transmission is the time it takes for users to set up a peer-to-peer connection. Submit handles peer-to-peer connections internally and automatically populates most settings before providing a unified user interface to the user; however, more in-depth design research including participatory design workshops and interviews would allow us to better understand how design a system that is both easy to use and also gives users control over data ownership and visibility.

8.5.3 Conclusion

Both Submit and VillageShare provide solutions that assist with content creation and data collection from mobile applications operating in challenged environments where connectivity is non-existent or lower-capacity than in the typical use cases for which the applications were designed.

Submit improves deployments of data collection efforts in challenged environments by supporting variations in deployment contexts in a systematic manner. Submit's abstractions decouple data and connectivity to enable application-level optimization of sparse heterogeneous networks. Submit identifies available connectivity in challenged networking environments and sends appropriate data over available channels in limited resources

settings. Creating software tools that enable application-level communication optimizations through the selection of appropriate data for transmission over available network channels represents a necessary complement to infrastructural improvement. Data communication needs to be adaptable to deployment conditions and solu- tions that focus on optimizations to the network transport layer do not have the flexibility to leverage the sparse challenged network conditions that exist. Therefore, adaptable frameworks that create abstractions that target application-level users (as opposed to developers adapting to network transport layers) are needed to empower deployment architects to easily customize application deployments to match an organizations requirements. For mobile tools to be successful in resource-constrained environments they should be composable by non-programmers, deployable by resource-constrained organizations, usable by minimally trained users, and robust to intermittent power and networking outages.

VillageShare empowers rural students and schools to create content offline and coordinate collaborations between multiple users who are connecting from disconnected or poorly connected regions. While VillageShare for schools has not yet been deployed, its predecessor (Kwaabana) has been deployed as a single rural server architecture in Macha, Zambia [108]. Based on this deployment in addition to our own anticipations about a deployment specific to rural schools, we identify several usage questions we would like to explore with data from a deployment. These include questions pertaining to the frequency, scope, and content types associated with file sharing. With a more focused understanding on how rural school children use a file sharing system, we gain better insight into how future efforts towards broadband development might best serve these communities. In particular, because VillageShare uses explicit sharing and privileges the sharing of locally created content, it provides special insight into the development of localized knowledge systems. Data from these deployments have the potential to transform our understanding about information creation and mobilization in developing contexts.

We believe that VillageShare provides a complementary solution to systems that focus on improving access to information via the World Wide Web. We hope that in doing so, VillageShare will act as an aid in creating effective knowledge systems that integrate both global and indigenous information and improve the impact of technologies for rural education.

8.6 Acknowledgements

Submit was designed and evaluated in collaboration with Waylon Brunette, Gaetano Borriello, Fahad Pervaiz, Shahar Levari, and Richard Anderson. Submit also benefited from the feedback of Elizabeth Belding and Neha Kumar. We thank Samuel Sudar for his help on the evaluation of Submit. Work on Submit was supported by NSF research grant IIS-1111433, NSF Graduate Research Fellowship grants DGE-0718124 and DGE-1144085, and USAID contract AID-OAAA-13-00002.

VillageShare was designed and evaluated in collaboration with David Johnson, Paul Schmitt, and Elizabeth Belding. Thanks are due to the Meraka Institute at the Council for Scientific and Industrial Research in South Africa for their assistance in understanding the challenges associated the design of content synchronization systems in rural and developing contexts. VillageShare was supported by NSF Network Science and Engineering (NetSE) Award CNS-1064821 and NSF Graduate Research Fellowship Grant No. DGE-1144085.

Chapter 9

Conclusion and Future Directions

9.1 Conclusion

Access to the Internet is critical to the information capital of all communities. Numerous studies have shown how Internet access increases to economic opportunity [192, 20, 121, 195]. For marginalized communities, Internet access plays a critical role in providing opportunities for political action [207, 115, 213, 167, 1, 75] and platforms for representation [56, 178, 41, 6]. For Indigenous communities, Internet access provides opportunities for knowledge systems to be preserved and revitalized and for development of cultural resilience [129, 76, 60, 178, 31, 6]. Unfortunately, access to the Internet is not universal and while efforts in infrastructural development have increased access rates in rural and developing communities, the divide has shifted rather than closed due to the increasingly size and responsiveness of Internet content and applications [91, 156, 121, 195]. In this dissertation, we argue that divides can be narrowed by first understanding how ICT infrastructures are used to serve community information needs and then enhancing those infrastructures to meet needs in ways that are more resource efficient and customizable to community needs and behaviors. We refer to the resulting network infrastructures

as "community-based networks," and our evaluations have demonstrated their ability to increase the value of infrastructure by increasing the number of community members whose information needs are served and by improving on the quality of services through more effective resource utilization.

This dissertation has made impact with respect to its intellectual contributions as well as its societal contributions. Work from this dissertation was published in premier computer science venues focusing on the World Wide Web and Computer Supported Cooperative Work and one publication received recognition as an honorable mention for best paper in the 2017 Conference for Computer Supported Collaborative Work [204]. In addition to being featured in research publication venues, our work was featured in a popular publication [72] and was delivered as a talk at the annual meeting of the National Congress of American Indians (NCAI) Technology Task Force Meeting in 2015. However, the most significant impact of this work is its focus on Internet connectivity issues that take place *in developed countries*. This was the first work to take an observational, network analytic approach to understanding connectivity issues in Indian Country and throughout the work, we have developed innovated mixed method approaches that enhance our understanding. We have collected a data set of over 115 TB worth of network packet headers and flow-level statistics generated as part of normal Internet usage in two tribal-operated networks that provide service to a total of 18 Native American reservations. While this data was not made public due to access limitations specified in the Internal Review Board (IRB) agreement we have with our collaborators, parts of the data set have been made available to networking classes at UC Santa Barbara.

9.2 Future Directions

We conclude with a brief discussion on future research directions made possible by this dissertation. We identify three main research categories, including: adaptive human-computer analysis systems, Internet of people, and community-centered protocols.

9.2.1 Adaptive human-computer analysis systems

With the rise of pervasive ICTs and paradigms such as the Internet of Things (IoT), data is generated at unprecedented volumes [94]. This leads to two significant challenges for effective data analysis. First, the majority of data is unstructured, meaning that data is generated in *ad hoc* formats that lack pre-determined structure or annotative tags. This lack of structure presents several significant challenges. Lack of structure makes the consistent application of analysis techniques over time virtually impossible, particularly as data forms evolve and co-exist in different formats. We faced this challenge in our own work (particularly since we collected network traces over such a long period of time). One example was in our analysis of Instagram traffic where multiple versions of the Instagram application were being used in the network at once, and each version presented a different form of meta data embedded in the URL header for the same API call. Accurate analysis required that we characterize the variety present in the network for a single Instagram API call and then deconstruct the structure of each varietal. This approach obviously inhibits automatic and real-time streaming analysis of some of the data that is most meaningful to users, which in turn prevents systems from being truly intelligent and responsive to the changing needs and behaviors of users. A second challenge is the lack of data veracity, meaning data lacks guarantees about semantic value. With such a large volume of data available, it is difficult to know which data streams are most integral to answering critical questions and which streams lack semantic or analytic value. This can

lead to expensive storage and analysis resources being dedicated to data that produces no or low-quality insights.

The unique combination of methodologies and findings in this dissertation lead us to believe that we can shed some new insights into these challenges as they pertain to the creation of sustainable information systems for communities. First, our work and the work of others demonstrates that aggregate usage behavior changes over time in response to infrastructure capacity [216, 39], application capabilities [216, 200, 204, 211], and information relevance [204]. In order to create systems of analysis that can appropriately adapt to these changes in behavior, we suggest investigation into hybridized data analysis systems that integrate the resource efficiency of automated analysis techniques with the wisdom of human insight and direction. For example, imagine if a community decided to deploy our FM radio content distribution system to distribute the most relevant content (as perceived in aggregate) to their community [200]. To establish which content was most relevant (which is a very value laden term), there would need to be some qualitative and ethnographic analysis performed to answer questions such as: *What kind of information is most important to you? What are characteristics of information sources you trust?*. The results of these analysis would be used to first identify relevant information sources (e.g., the usage traces of others or content from specific Web domains) and then filter large, noisy data streams in accordance with the features identified by users. As content is collected over time via automatic processes, these techniques could be monitored by auxiliary processes. These auxiliary processes would monitor the meta-data indicative of significant changes to data structure or semantics that might alter the value of data that is being collected to determine relevancy. Additionally, these auxiliary processes would alert human analysts in the loop to a potentially significant change in the underlying value of data or content collected with aggregated examples and summary meta-data. Human analysts might affirm these alerts and respond with manual alterations of the

automatic data collection and analysis processes (e.g., add support for the collection of data from an application with an updated API or remove a source of content from the list of "relevant and trusted" information sources). Human analysts might also indicate alerts as false positive indicators of significant change, giving these auxiliary analysis processes the opportunity to adapt to new information about what constitutes substantive change.

Research into adaptive human-computer analysis systems would be important for information systems operating in resource-constrained contexts, but it would have significant impact on the larger global economy as well. As more jobs are automated and as an increasing amount of data is being generated, there is a need both for human work opportunities and human analysts [163, 93]. Adaptive human-computer analysis refines large, challenging analysis jobs to smaller, more specialized analysis tasks that can ultimately lead to greater semantic value of insights resulting from the analysis.

9.2.2 Internet of people

The Internet of Things (IoT) refers to "a dynamic global network infrastructure with self configuring capabilities based on standard and interoperable communication protocols where physical and virtual 'things' have identities, physical attributes, and virtual personalities, use intelligent interfaces, and are seamlessly integrated into the information network [198]." Unfortunately, the implementation of IoT has left much to be desired, requiring users to maintain constant awareness of system interactions and limitations [128]. A relatively new and counter-balancing paradigm to IoT is the Internet of People (IoP) [128, 28, 117], which seeks to augment IoT to ensure that devices are serving the information needs of people, rather than requiring human users to have "slavish awareness" to devices and systems [128]. IoP is also a helpful construct for thinking about the semantics of Internet applications. For example, interactions that

take place across multiple social media platforms fundamentally constitute interactions between people.

Insights from this dissertation point to two veins of research that leverage the Internet of People paradigm. First is the need for software and platforms that integrate devices for resource optimization [26]. Our work with ODK Submit in Chapter 8.3 demonstrated how a platform interface on a single device could ease the burden of navigating connectivity context and network resource allocation. Now, as users have multiple devices with many different types of network interfaces, there is a need to coordinate these devices to a single user in such a way that a user's network resources are not overwhelmed by competing devices and a user's information needs are met appropriately. This requires mechanisms wherein devices are tied to individuals so that they can coordinate resource usage in meaningful ways. This leads to the need for new thinking on privacy and security guarantees as well as new abstractions that give individuals transparency into the data they generate and steps they can take to provide more control over that data.

Second is the need for mechanisms that allow people to share information with individuals and communities using abstractions that simplify this process while also providing people with more control over the semantics of information flow [204]. For example, imagine a platform that seeks to connect individuals to each other, allows those individuals to be labeled as belonging to certain communities (through self-selection, collaborative voting processes, or assignment), and allows users to specify formats and channels on which they would like to receive specific topics of information from other people. For example, a user might only want to receive "Work" related information at a specific email address and they might want to receive "Local news" information on a certain social media platform. This type of personal interface would also provide abstractions for users to disseminate information more effectively as well. For instance, they could generate content intended for a community and simply by labeling the content with a

general topic, urgency level, and intended audience, the platform could manage where that content was posted and could provide automatic annotations (e.g., tagging) that would optimize rates of diffusion and prevalence.

9.2.3 Community-centered protocols

In this dissertation, we introduce the ways that community usage patterns and preferences can be leveraged to add value to existing infrastructures [203, 200]. However, in designing these community-centered systems, there are needs for community-centered protocols. For example, existing protocols for data ownership and infrastructure governance exist for individuals and corporations that function as individuals from a protocol standpoint. However, these protocols are insufficient when considering the expansion of community-based networks and the group-based usage patterns for which technologists often do not design [192].

We identify two specific research directions within the area of community-centered protocol design. First, there are issues that pertain to community infrastructures and the ensuing data themselves. Research questions include: *Who decides the information objectives of a community? Who owns (controls) content generated for a community content platform? Who is responsible for managing and maintaining community-based networks?* These questions represent protocol issues at a very high level, but the implications and answers have far-reaching consequences for the continued development of community-centered technologies.

The second direction for research involves community-centered protocols that reconciles the fact that communities are made of individuals. A specific example is mentioned in Chapter 7, which points to the fact that there are currently no existing mechanisms whereby an individual can entrust content access certificates to others with specific pri-

vacy and security guarantees. Developing such mechanisms would require research into a synthesis of security in distributed systems as well as a deeper understanding of human trust circles as pertains to information exchanges that take place between individuals and between individuals and organizations. A related direction for future research might focus on the proper definition of community from a trust perspective. While current methodologies (including those used in this dissertation [204, 200, 203]) perform community detection with graph-based analysis, these techniques represent a *post hoc* method for community definition. In real-time, these approaches lack the dynamism required to maintain up-to-date trust guarantees and they do not semantically convey trusted community as would be necessary for the exchange of sensitive information. Thus, we recommend investigation into alternative modes of community definition, including protocols that involve community assignment (which relates back to the need for community governance protocols), self-selection into communities, and membership voting schemes.

List of Terms

American Indian "American Indian is now a legal term that emerged out of common use by colonial authorities and settlers who, since the late 1500s, were erroneously describing the original indigenous inhabitants of what are now the Americas as 'indios,' or 'Indians.' The term 'American Indian' is used in many of the treaty documents negotiated between tribal peoples and U.S. colonial authorities, even though tribal peoples continue to recognize themselves as a people by the names of their tribe, i.e. Navajo or Diné, for Navajo Nation, and not according to the generalized population of indigenous peoples of the Americas or the English language term 'American Indian,' as neither of these articulates the inherent sovereign rights of tribes [58]." See also **Native American**.

BIA Bureau of Indian Affairs.

challenged environment A context that has limited access to infrastructural resources due to factors such as rurality, poverty, terrain, or policy.

cloud "Clouds are a large pool of easily usable and accessible virtualized resources (such as hardware, development platforms and/or services). These resources can be dynamically reconfigured to adjust to a variable load (scale), allowing also for an optimum resource utilization. This pool of resources is typically exploited by a

pay-per-use model in which guarantees are offered by the Infrastructure Provider by means of customized [service level agreements] [196]".

CSCW Computer Supported Cooperative Work.

FCC Federal Communications Commission.

ICT Information and Communication Technologies.

ICT4D Information and Communication Technologies for Development.

Indian Country Indian Country is a legal term that refers to the federally-recognized tribes, state-recognized tribes, pueblos, rancherias, bands, and Alaska Native villages and corporations within the political boundaries of the U.S. Used colloquially and not in a legal sense whatsoever, Indian Country also refers to Native peoples habits and norms in this somewhat parallel society. As a legal term, the phrase Indian Country has come to have meaning out of the basis of over a century of treaty-making and recognition processes between Native peoples and U.S. federal authorities. It inherently refers to an intertribal state of being for Native peoples in the U.S. [58].

IRB Internal Review Board.

ISP Internet Service Provider.

ITU International Telecommunications Union.

Native A less formal abbreviation of "Native American." This term can also be used as a more inclusive reference to peoples around the world who are native to their homeland, not just peoples native to North America. See also **Native American**.

NCAI National Congress of American Indians.

SCTCA Southern California Tribal Chairmen's Association.

TDV Tribal Digital Village.

tribal sovereignty Tribes' authority to self-govern based on treaties made with federal and state governments and upheld by the U.S. Supreme Court, President, and Congress. Self-governance includes hunting and fishing rights both within and without the boundaries of Indian Country, the authority to define tribal membership, manage tribal property, and regulate tribal business and domestic relations [136].

tribe Refers to historic and extant clans, tribes, bands, and nations formed of Native American people. It is critical to note that not all tribes are federally recognized..

WiFi Wireless, local area network products based on the 802.11 standard.

WISP Wireless Internet Service Provider.

Bibliography

[1] S. Agarwal, W. Bennett, C. Johnson, and S. Walker. A Model of Crowd Enabled Organization: Theory and Methods for Understanding the Role of Twitter in the Occupy Protests. *International Journal of Communication*, 8:646–672, 2014.

[2] T. Ahtone. Radio on the Reservation. `http://projects.aljazeera.com/2014/reservation-radio/index.html`, March 2014.

[3] T. Ahtone. How media did and did not report on Standing Rock: Native American issues are only media sexy when natives with painted faces and horses are around. `http://www.aljazeera.com/indepth/opinion/2016/12/media-report-standing-rock-161214101627199.html`, December 2016.

[4] R. K. Q. Akee and J. B. Taylor. *Social and Economic Change on American Indian Reservations: A Databook of the US Censuses and the American Community Survey 1990-2010*. The Taylor Policy Group, Inc., Sarasota, FL, USA, May 2014.

[5] Alexa. Top Sites in United States. `http://www.alexa.com/topsites/countries/US`, October 2014.

[6] V. Alia. The New Media Nation: Indigenous Peoples and Global Communication. In *Anthropology of Media*. Berghan Books, New York, NY, USA, 2012.

[7] V. Anantraman, T. Mikkelsen, R. Khilnani, V. S. Kumar, R. Machiraju, A. Pentland, and L. Ohno-Machado. Handheld Computers for Rural Healthcare, Experiences in a Large Scale Implementation. In *Development by Design Conference Proceedings*, 2002.

[8] Apache Stronghold. Once Again, the Fight for Religious Freedom in America Begins. `http://www.apache-stronghold.com/about.html`, February 2016.

[9] 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.

[10] D. J. Aron and D. E. Burnstein. Broadband Adoption in the United States: An Empirical Analysis. In A. L. Shampine, editor, *Down to the Wire: Studies in the Diffusion and Regulation of Telecommunications Technologies*. Nova Publishers, 2003.

[11] P. Bahl, A. Adya, J. Padhye, and A. Walman. Reconsidering Wireless Systems with Multiple Radios. *SIGCOMM Computing Communication Review*, 34(5):39–46, October 2004.

[12] A. Balasubramanian, R. Mahajan, and A. Venkataramani. Augmenting Mobile 3G Using WiFi. In *Proc of the 8th International Conference on Mobile Systems, Applications, and Services*, MobiSys '10, pages 209–222, 2010.

[13] 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.

[14] C. Baldy. The New Native Intellectualism:#ElizabethCook-Lynn, Social Media Movements, and the Millennial Native American Studies Scholar. *Wicazo Sa Review*, 31(1):90–110, 2016.

[15] W. Bennett and A. Segerberg. The Logic of Connective Action. *Information, Communication & Society*, 15(5):739–768, 2012.

[16] B. Bimber. *Information and American Democracy: Technology in the Evolution of Political Power*. Cambridge University Press, 2003.

[17] V. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre. Fast Unfolding of Communities in Large Networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[18] C. J. Bonk. How Technology is Changing School. *Educational Leadership*, 67(7):60–65, 2010.

[19] d. boyd, S. Golder, and G. Lotan. Tweet, Tweet, and Retweet: Conversational Aspects of Retweeting on Twitter. In *HICSS-43*, Kauai, HI, USA, January 2010.

[20] E. Brewer, M. Demmer, B. Du, M. Ho, M. Kam, S. Nedevschi, J. Pal, R. Patra, S. Surana, and K. Fall. The Case for Technology in Developing Regions. *Computer*, 38(6):25–38, 2005.

[21] 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.

[22] J. Briggs. The Use of Indigenous Knowledge in Development: Problems and Challenges. *Progress in Development Studies*, 5(2):99–114, 2005.

[23] Bro Project. Bro Network Security Monitor. https://www.bro.org/, October 2014.

[24] 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.

[25] W. Brunette, M. Sundt, N. Dell, R. Chaudhri, N. Breit, and G. Borriello. Open Data Kit 2.0: Expanding and Refining Information Services for Developing Regions. In *Proc of the 14th Workshop on Mobile Computing Systems & Applications*, HotMobile '13, 2013.

[26] W. Brunette, M. Vigil, F. Pervaiz, S. Levari, G. Borriello, and R. Anderson. Optimizing Mobile Application Communication for Challenged Network Environments. In *Proceedings of the 2015 Annual Symposium on Computing for Development*, pages 167–175, 2015.

[27] Bureau of Indian Affairs. Who We Are. https://www.bia.gov/WhoWeAre/index.htm, April 2017.

[28] L. Butgereit and L. Coetzee. Beachcomber: Linking the "Internet of Things" to the "Internet of people". In *IST-Africa Conference Proceedings*, pages 1–7, Gaborone, Botswana, May 2011.

[29] D. Camps-Mur, A. Garcia-Saavedra, and P. Serrano. Device-to-Device Communications with Wi-Fi Direct: Overview and Experimentation. *Wireless Communications, IEEE*, 20(3):96–104, June 2013.

[30] I. Caputo. Solar Power Makes Electricity More Accessible On Navajo Reservation. http://www.npr.org/2015/04/21/401000427/solar-power-makes-electricity-more-accessible-on-navajo-reservation, August 2015.

[31] P. Carpenter, K. Gibson, C. Kakekaspan, and S. ODonnell. How Women in Remote and Rural First Nation Communities are Using Information and Communication Technologies (ICT). *Journal of Rural and Community Development*, 8(2):79–97, 2014.

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

[33] P. Casas, A. Sackl, S. Egger, and R. Schatz. YouTube & Facebook Quality of Experience in Mobile Broadband Networks. In *Globecom Workshops (GC Wkshps), 2012 IEEE*, pages 1269–1274. IEEE, 2012.

[34] M. Castells. *The Information Age: Economy, Society and Culture*, volume 2. Blackwell Press, Malden, MA, USA, 1997.

[35] Center for Native American Youth at the Aspen Institute. Fast Facts on Native American Youth and Indian Country. `http://www.aspeninstitute.org/sites/default/files/content/images/Fast%20Facts.pdf`, April 2011.

[36] Central Intelligence Agency. World Fact Book. `https://www.cia.gov/Library/publications/the-world-factbook/fields/2213.html`, June 2014.

[37] D. Chafekar and H. Armstrong. State of Twitter. `https://www.quettra.com/research/state-of-twitter/`, October 2015.

[38] J. Chen, D. Hutchful, W. Thies, and L. Subramanian. Analyzing and Accelerating Web Access in a School in Peri-urban India. In *Proceedings of the 20th International Conference Companion on World Wide Web*, pages 443–452. ACM, 2011.

[39] J. Chen, L. Subramanian, and J. Li. RuralCafe: Web Search in the Rural Developing World. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 411–420, Madrid, Spain, May 2009.

[40] 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.

[41] M. Clark. *To Tweet Our Own Cause: A Mixed-Methods Analysis of the Online Phenomena Known as Black Twitter.* PhD thesis, University of North Carolina at Chapel Hill, 2014.

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

[43] Common Core State Standards Initiative. Read the Standards. `http://www.corestandards.org/read-the-standards/`, 2017.

[44] G. Coulthard. *Red Skin, White Masks: Rejecting the Colonial Politics of Recognition.* University of Minnesota Press, Minneapolis, MN, USA, 2014.

[45] R. Daft and R. Lengel. Information Richness: A New Approach to Manager Information Processing and Organisational Design, 1984.

[46] G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, and W. Vogels. Dynamo: Amazon's Highly Available Key-Value Store. In *ACM SIGOPS Operating Systems Review*, number 6, pages 205–220, 2007.

[47] L. B. Deek, K. C. Almeroth, M. P. Wittie, and K. A. Harras. Exploiting Parallel Networks Using Dynamic Channel Scheduling. In *Proc of the 4th Annual International Conference on Wireless Internet*, WICON '08, pages 1–9, ICST, Brussels, Belgium, Belgium, 2008.

[48] M. J. Demmer, B. Du, and E. A. Brewer. TierStore: A Distributed Filesystem for Challenged Networks in Developing Regions. In *Proceedings of FAST*, volume 8, pages 1–14, 2008.

[49] S. Deng, A. Sivaraman, and H. Balakrishnan. All Your Network Are Belong to Us: A Transport Framework for Mobile Network Selection. In *Proc of the 15th Workshop on Mobile Computing Systems & Applications*, HotMobile '14, 2014.

[50] Denso Wave Incorporated. QR Code. `http://www.qrcode.com/en/about/version.html`.

[51] 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.

[52] B. DeRenzi, N. Lesh, T. Parikh, C. Sims, W. Maokla, M. Chemba, Y. Hamisi, M. Mitchell, G. Borriello, et al. E-IMCI: Improving Pediatric Health Care in Low-income Countries. In *Proceedings of the Conference on Human Factors in Computing Systems*, CHI '08, pages 753–762, 2008.

[53] B. DeRenzi, C. Sims, J. Jackson, G. Borriello, and N. Lesh. A Framework for Case-Based Community Health Information Systems. In *Global Humanitarian Technology Conference (GHTC)*, pages 377–382. IEEE, 2011.

[54] 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.

[55] T. Dreher, K. McCallum, and L. Waller. Indigenous Voices and Mediatized Policymaking in the Digital Age. *Information, Communication & Society*, 19(1):23–39, 2016.

[56] E. Dreyfuss. As Standing Rock Protesters Face Down Armored Trucks, the World Watches on Facebook. `https://www.wired.com/2016/10/standing-rock-protesters-face-police-world-watches-facebook/`, October 2016.

[57] E. Dreyfuss. Social Media Made the World Care About Standing Rockand Helped It Forget. `https://www.wired.com/2017/01/social-media-made-world-care-standing-rock-helped-forget/`, January 2017.

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

[59] M. Duggan, N. B. Ellison, C. Lampe, A. Lenhart, and M. Madden. Social Media Update 2014. http://www.pewinternet.org/2015/01/09/social-media-update-2014/, January 2015.

[60] R. Dunbar-Ortiz. *An Indigenous Peoples' History of the United States*, volume 3. Beacon Press, 2014.

[61] W. Echo-Hawk. *In the Courts of the Conqueror: The Ten Worst Indian Law Cases Ever Decided*. Fulcrum Publishing, 2010.

[62] Energy Information Administration. Energy Consumption and Renewable Energy Development Potential on Indian Lands. https://www.eia.gov/renewable/archive/neaf0001.pdf, April 2010.

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

[64] K. Fall. A Delay-tolerant Network Architecture for Challenged Internets. In *Proceedings of the 2003 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, pages 27–34. ACM, 2003.

[65] Federal Communications Commission. Availability. In *National Broadband Plan: Connecting America*. January 2010.

[66] Federal Communications Commission. Mobile Broadband: The Benefits of Additional Spectrum. https://transition.fcc.gov/national-broadband-plan/mobile-broadband-paper.pdf, October 2010.

[67] Federal Communications Commission. Taking Additional Steps to Make U.S. Spectrum Policy More Comprehensive. http://www.broadband.gov/plan/5-spectrum/#s5-3, March 2010.

[68] Federal Communications Commission. 2012 annual report, March 2013.

[69] Federal Communications Commission. Tribal Initiatives. http://transition.fcc.gov/indians, March 2013.

[70] Federal Communications Commission. Native Nations Consultation and Policy. http://www.fcc.gov/encyclopedia/native-nations-consultation-and-policy, December 2014.

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

[72] S. Fernandez. Internet in Wide Open Spaces. http://www.news.ucsb.edu/2016/017365/internet-wide-open-spaces?utm_source=newsletter&utm_medium=email&utm_content=Read%20more&utm_campaign=November%203%2C%202016, November 2016.

[73] R. Gandhi, R. Veeraraghavan, K. Toyama, and V. Ramprasad. Digital Green: Participatory Video for Agricultural Extension. In *International Conference on Information and Communication Technologies and Development*, pages 1–10. IEEE, 2007.

[74] K. Garret. Protest in an Information Society: A Review of Literature on Social Movements and New ICTs. *Information, Communication, & Society*, 9(2):202–224, 2006.

[75] M. Garrido and A. Halavais. Mapping Networks of Support for the Zapatista Movement. *Cyberactivism: Online activism in theory and practice*, pages 165–184, 2003.

[76] 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, Toronto, Canada, February 2012.

[77] 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, pages 1603–1612, Florence, Italy, 2008.

[78] P. Gloor. *Swarm Creativity: Competitive Advantage Through Collaborative Networks*. Oxford University Press, 2006.

[79] Goldman Environmental Prize. Homepage. http://www.goldmanprize.org, January 2016.

[80] A. Gonzales. The Contemporary U.S. Digital Divide: From Initial Access to Technology Maintenance. *Information, Communication & Society*, 19(2):234–248, 2016.

[81] Government Accountability Office. Tribal Internet Access: Increased Federal Coordination and Performance Measurement Needed: Statement of Mark Goldstein, Director, Physical Infrastructure, GAO-16-504T, April 2016.

[82] 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.

[83] K. Gush, G. Cambridge, and R. Smith. The Digital Doorway-Minimally Invasive Education in Africa. In *ICT in Education Conference*, volume 229, 2004.

[84] E. Guskin and A. Mitchell. Innovating News in Native Communities. `http://www.stateofthemedia.org/2012/native-american-news-media/`, January 2013.

[85] R. G. Guy, J. S. Heidemann, W.-K. Mak, T. W. Page Jr, G. J. Popek, D. Rothmeier, et al. Implementation of the Ficus Replicated File System. In *USENIX Summer*, pages 63–72, 1990.

[86] B. Han, P. Hui, V. A. Kumar, M. V. Marathe, J. Shao, and A. Srinivasan. Mobile Data Offloading Through Opportunistic Communications and Social Participation. *IEEE Transactions on Mobile Computing*, 11(5):821–834, May 2012.

[87] S. Harrison. May I See Your ID? How Voter Identification Laws Disenfranchise Native Americans' Fundamental Right to Vote. *American Indian Law Review*, 37(2):597–628, 2012.

[88] C. Hartung et al. Open Data Kit: Tools to Build Information Services for Developing Regions. In *Proc of the 4th ACM/IEEE Int Conf on Information and Communication Technologies and Development*, ICTD '10, 2010.

[89] R. Heeks. Avoiding eGov Failure: Design-Reality Gap Techniques. `www.egov4dev.org/success/techniques/drg.shtml`, 2008.

[90] K. Heimerl and E. Brewer. The Village Base Station. In *Proceedings of the 4th ACM Workshop on Networked Systems for Developing Regions*, page 14, 2010.

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

[92] Indian Country Today Media Network. Homepage. `http://indiancountrytodaymedianetwork.com/`, April 2016.

[93] International Business Machines. The 4 V's of Big Data. `http://www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-big-data.jpg`, May 2011.

[94] International Business Machines. What is Big Data? `https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html`, May 2011.

[95] International Telecommunications Union. The World in 2011: ICT Facts and Figures. `https://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2011-e.pdf`, 2011.

[96] International Telecommunications Union. The World in 2013: ICT Facts and Figures. https://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2013-e.pdf, 2013.

[97] International Telecommunications Union. Mobile Broadband, Smartphones, Apps, Fixed Networks. https://itunews.itu.int/En/4958-Mobile-broadband-smartphones-apps-fixed-networks.note.aspx, February 2014.

[98] International Telecommunications Union. ICT Facts and Figures 2015. http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2016.pdf, May 2015.

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

[100] Internet Live Stats. Internet Users by Country (2016). http://www.internetlivestats.com/internet-users-by-country/, May 2017.

[101] 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.

[102] P. Jaccard. The Distribution of the Flora in the Alpine Zone. *New Phytologist*, 11(2):37–50, 1912.

[103] V. Jacobson, C. Leres, and S. McCanne. tcpdump. http://www.tcpdump.org/tcpdump_man.html, Februrary 2017.

[104] R. Jain, D. Chiu, and W. Hawe. A Quantitative Measure of Fairness and Discrimination for Resource Allocation in Shared Computer Systems. *Technical Report, Digital Equipment Corporation, DEC-TR-301*, 1984.

[105] S. John. Idle No More-Indigenous Activism and Feminism. *Theory in Action*, 8(4):38–54, 2015.

[106] D. Johnson, V. Pejovic, E. Belding, and G. van Stam. Traffic Characterization and Internet Usage in Rural Africa. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 493–502, Hyderabad, India, March 2011.

[107] D. Johnson, V. Pejovic, E. Belding, and G. van Stam. VillageShare: Facilitating Content Generation and Sharing in Rural Networks. In *Proceedings of the 2nd ACM Symposium on Computing for Development*, ACM DEV '12, pages 61–70, Atlanta, GA, USA, March 2012.

[108] D. L. Johnson, E. M. Belding, and C. Mudenda. Kwaabana: File Sharing for Rural Networks. In *Proceedings of the 4th Annual Symposium on Computing for Development*, page 4, 2013.

[109] B. Keegan, S. Lev, and O. Arazy. Analyzing Organizational Routines in Online Knowledge Collaborations: A Case for Sequence Analysis in CSCW. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work & Social Computing*, March 2016.

[110] T. T. Keegan, P. Mato, and S. Ruru. Using Twitter in an Indigenous Language: An Analysis of Te Reo Maori Tweets. *AlterNative: An International Journal of Indigenous Peoples*, 11(1):45–58, 2015.

[111] J. J. Kistler and M. Satyanarayanan. Disconnected Operation in the Coda File System. *ACM Trans. Comput. Syst.*, 10(1):3–25, Feb 1992.

[112] A. Koenig and J. Stein. Federalism and the State Recognition of Native American Tribes: A Survey of State-recognized Tribes and State Recognition Processes Across the United States. *Santa Clara Law Review*, 48:79, 2008.

[113] D. Kravets. U.N. Report Declares Internet Access a Human Right. `https://www.wired.com/2011/06/internet-a-human-right/`, June 2011.

[114] J. F. Kurose and K. W. Ross. *Computer Networking: A Top-down Approach*, volume 7. Pearson, 2016.

[115] Lakota People's Law Project. Native Lives Matter. `http://www.docs.lakotalaw.org/reports/Native%20Lives%20Matter%20PDF.pdf`, February 2015.

[116] R. Lengel and R. Daft. An Exploratory Analysis of the Relationship Between Media Richness and Managerial Information Processing. Technical report, Texas A&M University, 1984.

[117] M. Li. Internet of People. *Concurrency and Computation: Practice and Experience*, 29(3), 2017.

[118] Y. Li, G. Su, P. Hui, D. Jin, L. Su, and L. Zeng. Multiple Mobile Data Offloading Through Delay Tolerant Networks. In *Proc of the 6th ACM Workshop on Challenged Networks*, CHANTS '11, pages 43–48, 2011.

[119] 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.

[120] S. Manning. Sanders Advocates for Tribes, Mother Earth. `http://indiancountrytodaymedianetwork.com/2016/05/23/manning-sanders-advocates-tribes-mother-earth-164564`, May 2016.

[121] S. P. Martin. Is the Digital Divide Really Closing? A Critique of Inequality Measurement in a Nation Online. *IT & society*, 1(4):1–13, 2003.

[122] M. Martins and R. Fonseca. Application Modes: A Narrow Interface for End-user Power Management in Mobile Devices. In *Proc of the 14th Workshop on Mobile Computing Systems and Applications*, HotMobile '13, 2013.

[123] 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.

[124] K. McCallum, M. Meadows, L. Waller, M. Dunne Breen, and H. Reid. *The Media and Indigenous Policy: How News Media Reporting and Mediatized Practice Impact on Indigenous Policy*. University of Canberra, 2012.

[125] D. McCaskill and J. Rutherford. Indigenous peoples of South-East Asia: poverty, identity and resistance. In R. Eversole, J.-A. McNeish, and A. D. Cimadamore, editors, *Indigenous Peoples and Poverty: An International Perspective*, pages 126–157. Zed Books, London, UK, December 2005.

[126] A. Mercado. Medios Indígenas Transnacionales: El Fomento del Cosmopolitismo Desde Abajo. *Comunicación y Sociedad*, 1(23):171–193, 2015.

[127] 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.

[128] J. Miranda, N. Mäkitalo, J. Garcia-Alonso, J. Berrocal, T. Mikkonen, C. Canal, and J. M. Murillo. From the Internet of Things to the Internet of People. *IEEE Internet Computing*, 19(2):40–47, 2015.

[129] 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):275–278, 2014.

[130] F. Morstatter, J. Pfeffer, H. Liu, and K. Carley. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In *ICWSM '13: Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, June 2013.

[131] S. Moya-Smith. Tara Houska, Ojibwe, Named Native American Advisor to Bernie Sanders. `http://indiancountrytodaymedianetwork.com/2016/02/24/tara-houska-ojibwe-named-native-american-advisor-bernie-sanders-163531`, Februrary 2016.

[132] J. Murphy and M. Roser. Internet. https://ourworldindata.org/internet/, January 2017.

[133] A. Muthitacharoen, B. Chen, and D. Mazières. A Low-bandwidth Network File System. *SIGOPS Operating Systems Review*, 35(5):174–187, Oct. 2001.

[134] MyMedLab. MyMedLab Story. http://www.mymedlab.com/about, 2010.

[135] D. Narayan-Parker. *Bonds and Bridges: Social Capital and Poverty*, volume 2167. World Bank Publications, 1999.

[136] National Congress of American Indians. Tribal Sovereignty. In *Tribal Nations and the United States: An Introduction*, pages 16–19. National Congress of American Indians, Washington, D.C., USA, 2014.

[137] National Parks Service. Indian Reservations in the Continental United States. https://www.nps.gov/nagpra/documents/resmap.htm, 2010.

[138] S. Nirjon, A. Nicoara, C.-H. Hsu, J. Singh, and J. Stankovic. MultiNets: Policy Oriented Real-Time Switching of Wireless Interfaces on Mobile Devices. In *Real-Time and Embedded Technology and Applications Symposium (RTAS), 2012 IEEE 18th*, pages 251–260, April 2012.

[139] Norris, Tina and Vines, Paula L. and Hoeffel, Elizabeth M. The American Indian and Alaska Native Population: 2010. https://www.census.gov/prod/cen2010/briefs/c2010br-10.pdf, January 2012.

[140] Obami. About Obami. http://www.obami.com/portals/obami/about_obami, January 2017.

[141] S. C. of the United States. Federal Power Commission v. Tuscarora Indian Nation, 1960.

[142] Office of Native Affairs and Policy. Native Nations. https://www.fcc.gov/general/native-nations, July 2016.

[143] Open Development Kit. Current Distribution. https://opendatakit.org/use/2_0_tools/active-distribution/, October 2016.

[144] Open Development Kit. Deployments. https://opendatakit.org/about/deployments/, October 2016.

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

[146] 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.

[147] ownCloud. ownCloud. `http://owncloud.org/`, February 2014.

[148] O. Oyman and S. Singh. Quality of Experience for HTTP Adaptive Streaming Services. *IEEE Communications Magazine*, 50(4):44–51, 2012.

[149] N. D. Parkhurst, T. Morris, E. Tahy, and K. Mossberger. The Digital Reality: E-government and Access to Technology and Internet for American Indian and Alaska Native Populations. In *Proceedings of the 16th Annual International Conference on Digital Government Research*, pages 217–229, 2015.

[150] R. K. Patra, S. Nedevschi, S. Surana, A. Sheth, L. Subramanian, and E. A. Brewer. WiLDNet: Design and Implementation of High Performance WiFi Based Long Distance Networks. In *NSDI*, pages 87–100, 2007.

[151] V. Paxson. End-to-end Routing Behavior in the Internet. *IEEE/ACM Transactions on Networking*, 5(5):601–615, 1997.

[152] 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.

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

[154] T. Perrelli. Statement of Associate Attorney General Perrelli before the Committee on Indian Affairs on Violence Against Native American Women [citing a National Institute of Justice-funded analysis of death certificates]. `http://www.justice.gov/iso/opa/asg/speeches/2011/asg-speech-110714.html`, July 2011.

[155] S. Pevar. *The Rights of Indians and Tribes*. Oxford University Press, 2012.

[156] Pew Research Center. World Wide Web Timeline. `http://www.pewinternet.org/2014/03/11/world-wide-web-timeline/`, March 2014.

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

[158] Politecnico di Torino. Ttat: TCP Statistic and Analysis Tool. `http://tstat.polito.it/`, May 2016.

[159] Z. Quan, S. Cui, and A. H. Sayed. Optimal Linear Cooperation for Spectrum Sensing in Cognitive Radio Networks. *IEEE Journal of Selected Topics in Signal Processing*, 2(1):28–40, 2008.

[160] A. Rabagliati. Wizzy Digital Courier–How it Works. `http://wizzy.org.za`, April 2004.

[161] A. Rahmati, L. Zhong, V. Vasudevan, J. Wickramasuriya, and D. Stewart. Enabling Pervasive Mobile Applications with the FM Radio Broadcast Data System. In *Proceedings of the Eleventh Workshop on Mobile Computing Systems and Applications*, HotMobile '10, pages 78–83, Annapolis, MD, USA, February 2010.

[162] Red Spectrum Communications. Red Spectrum Services. `http://www.red-spectrum.com/services.html`, January 2010.

[163] J. Rifkin. *End of Work*. Pacifica Radio Archives, 1996.

[164] S. Rogers. What fuels a Tweets engagement? `https://blog.twitter.com/2014/what-fuels-a-tweets-engagement`, October 2014.

[165] Safaricom. M-pesa. `https://www.safaricom.co.ke/personal/m-pesa`, May 2014.

[166] C. Sandvig. Connection Up Ewiiaapaayp Mountain. In L. Nakamura and P. Chow-White, editors, *Race after the Internet*, pages 168–200. Routledge, New York, NY, USA, 2013.

[167] S. Savage and A. Monroy-Hernández. Participatory Militias: An Analysis of an Armed Movement's Online Audience. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 724–733, March 2015.

[168] M. Schandorf and A. Karatzogianni. #NODAPL: Distributed Rhetorical Praxis at Standing Rock. In *Digital Writing & Rhetoric*. Routeledge, 2017.

[169] R. Schatz, S. Egger, and A. Platzer. Poor, Good Enough or Even Better? Bridging the Gap between Acceptability and QoE of Mobile Broadband Data Services. In *2011 IEEE International Conference on Communications*, pages 1–6, June 2011.

[170] J. Schroedel and A. Aslanian. Native American Vote Suppression: The Case of South Dakota. *Race, Gender & Class*, 22(1/2):308, 2015.

[171] J. Scott et al. Haggle: A Networking Architecture Designed Around Mobile Users. In *WONS 2006: Third Annual Conference on Wireless On-demand Network Systems and Services*, pages 78–86, 2006.

[172] A. Seth, D. Kroeker, M. Zaharia, S. Guo, and S. Keshav. Low-cost Communication for Rural Internet Kiosks Using Mechanical Backhaul. In *Proceedings of the 12th Annual International Conference on Mobile Computing and Networking*, MobiCom '06, pages 334–345, New York, NY, USA, 2006. ACM.

[173] S. Shah and A. Joshi. COCO: A Web-based Data Tracking Architecture for Challenged Network Environments. In *Proceedings of the First ACM Symposium on Computing for Development*, ACM DEV '10, pages 7:1–7:7, New York, NY, USA, 2010. ACM.

[174] K. Siegler. Radio Station KYAY is Lifeline For Apache Tribe. `http://www.npr.org/2013/09/03/218455207/radio-station-kyay-is-lifeline-for-apache-tribe`, September 2013.

[175] Siemens. Siemens to supply power to Native American reservation in California. `http://www.siemens.com/press/en/pressrelease/?press=/en/pressrelease/2015/energymanagement/pr2015080313emen.htm&content[]=EM`, August 2015.

[176] A. Smith. Why Americans Use Social Media. `http://www.pewinternet.org/2015/01/09/social-media-update-2014/`, November 2011.

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

[178] L. T. Smith. *Decolonizing Methodologies: Research and Indigenous Peoples*. Zed Books, London, 2012.

[179] Southern California Tribal Digital Village. About TDV. `https://sctdv.net/about-tdv/`, January 2017.

[180] R. W. Stevens and G. R. Wright. TCP/IP Illustrated: Vol. 2: The Implementation, 1995.

[181] Substance Abuse and Mental Health Services Administration. American Indian and Alaska Native Culture Card: A Guide to Build Cultural Awareness. `https://store.samhsa.gov/product/American-Indian-and-Alaska-Native-Culture-Card/SMA08-4354`, March 2009.

[182] D. B. Terry, M. M. Theimer, K. Petersen, A. J. Demers, M. J. Spreitzer, and C. H. Hauser. Managing Update Conflicts in Bayou, a Weakly Connected Replicated Storage System. In *Proceedings of the Fifteenth ACM Symposium on Operating Systems Principles*, SOSP '95, pages 172–182, Copper Mountain, Colorado, USA, 1995.

[183] P. Tjaden and N. Thoennes. Full Report of the Prevalence, Incidence, and Consequences of Violence Against Women: Findings from the National Violence Against Women Survey. http://www.ncjrs.gov/txtfiles1/nij/183781.txt, November 2000.

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

[185] 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.

[186] Z. Tufekci. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. In *ICWSM '14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, May 2014.

[187] J. Tully. *Strange Multiplicity: Constitutionalism in an Age of Diversity*, volume 1. Cambridge University Press, 1995.

[188] J. Tully. *Public Philosophy in a New Key*, volume 1. Cambridge University Press Cambridge, 2008.

[189] J. Tveten. On American Indian reservations, challenges perpetuate the digital divide. https://arstechnica.com/information-technology/2016/01/on-american-indian-reservations-challenges-perpetuate-the-digital-divide/, January 2016.

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

[191] University of California, Berkeley Library. Standing Rock and the Dakota Access Pipeline: Native American Perspectives: Social Media. http://guides.lib.berkeley.edu/c.php?g=585158&p=4043507, January 2017.

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

[193] U.S. Department of Commerce and the National Telecommunications and Information Administration. Exploring the Digital Nation: Embracing the Mobile Movement. http://www.ntia.doc.gov/files/ntia/publications/exploring_the_digital_nation_embracing_the_mobile_internet_10162014.pdf, October 2014.

[194] M. Van Der Velden. Knowledge Facts, Knowledge Fiction: The Role of ICTs in Knowledge Management for Development. *Journal of International Development*, 14(1):25–37, 2002.

[195] J. A. Van Dijk. Digital Divide Research, Achievements and Shortcomings. *Poetics*, 34(4-5):221–235, 2006.

[196] L. M. Vaquero, L. Rodero-Merino, J. Caceres, and M. Lindner. A Break in the Clouds: Towards a Cloud Definition. *ACM SIGCOMM Computer Communication Review*, 39(1):50–55, 2008.

[197] C. Velthuis, P. Fisser, and J. Pieters. Collaborative Curriculum Design to Increase Science Teaching Self-Efficacy: A Case Study. *Journal of Educational Research*, 108(3):217–225, 2015.

[198] O. Vermesan, P. Friess, P. Guillemin, S. Gusmeroli, H. Sundmaeker, A. Bassi, I. S. Jubert, M. Mazura, M. Harrison, M. Eisenhauer, et al. Internet of Things Strategic Research Roadmap. *Internet of Things-Global Technological and Societal Trends*, 1:9–52, 2011.

[199] 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.

[200] 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.

[201] M. Vigil, D. Johnson, and E. Belding. Poster: Localized Content for Village Schools. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*, pages 491–491, 2015.

[202] 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.

[203] M. Vigil-Hayes, E. Belding, and E. Zegura. FiDO: A Community-based Web Browsing Agent and CDN for Challenged Environments. In *in submission*, May 2017.

[204] 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.

[205] W. W. Vithanage and A. S. Atukorale. Bassa: A Time Shifted Web Caching System for Developing Regions. In *Proceedings of the 5th ACM Workshop on Networked Systems for Developing Regions*, pages 63–68, 2011.

[206] M. Vitos, J. Lewis, M. Stevens, and M. Haklay. Making Local Knowledge Matter: Supporting Non-literate People to Monitor Poaching in Congo. In *Proc of the 3rd ACM Symp on Computing for Development*, ACM DEV '13, 2013.

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

[208] 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.

[209] D. Watts and S. Strogatz. Collective Dynamics of 'Small-world' Networks. *Nature*, 393(6684):440–442, 1998.

[210] L. Wei-Chih et al. UjU: SMS-based Applications Made Easy. In *Proc of the First ACM Symposium on Computing for Development*, ACM DEV '10, 2010.

[211] L. Wei-Chih, M. Tierney, J. Chen, F. Kazi, A. Hubard, J. G. Pasquel, L. Subramanian, and B. Rao. UjU: SMS-based Applications Made Easy. In *Proceedings of the First ACM Symposium on Computing for Development*, pages 166–177, 2010.

[212] J. Wolfley. You Gotta Fight for the Right to Vote: Enfranchising Native American Voters. *University of Pennsylvania Journal of Constitutional Law*, 18(1), 2015.

[213] V. Wulf, K. Aal, I. Abu Kteish, M. Atam, K. Schubert, M. Rohde, G. Yerousis, and D. Randall. Fighting Against the Wall: Social Media Use by Political Activists in a Palestinian Village. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1979–1988, 2013.

[214] K.-K. Yap et al. Making Use of All the Networks Around Us: A Case Study in Android. In *Proc of the 2012 ACM SIGCOMM Workshop on Cellular Networks: Operations, Challenges, and Future Design*, CellNet '12, pages 19–24, 2012.

[215] 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.

[216] M. Zheleva, P. Schmitt, M. Vigil, and E. Belding. Internet Bandwidth Upgrade: Implications on Performance and Usage in Rural Zambia. *Information Technologies & International Development*, 11(2):1–17, 2015.

[217] W. Zhou, E. Simpson, and D. P. Domizi. Google Docs in an Out-of-Class Collaborative Writing Activity. *International Journal of Teaching and Learning in Higher Education*, 24(3):359–375, 2012.