Abstract

Cellular service providers are under tremendous pressure to accommodate the burgeoning data needs of mobile subscribers. One intriguing solution is leveraging co-location, namely taking advantage of proximity between mobile devices for opportunistic collaboration. Co-location affords the usage of short range, low-power radio interfaces, avoiding expensive cellular interfaces in orthogonal device-to-device (D2D) channels. Taken one step further, could mobile devices exchange caches when in proximity, pre-staging popular content with one another? A first step to such research is understanding the potential with actual users. In this paper, we analyze our NetSense study of nearly two hundred smartphone users to explore the potential for pre-staging. We find that not only is the propagation velocity of data via co-location exchanges reasonable, the actual storage costs for pre-staging are also quite reasonable and within the normal free space of typical smartphone users.

1. Introduction

The wireless service industry, particularly cellular service providers, face overwhelming demand challenges over the next few years. Users are deploying new mixes of ever more capable devices while the available content and demand for said content on mobile devices grows unabated. The net result is a scramble by service providers to meet such needs taking an all of the above strategy, embracing techniques for capacity growth (small cells, spectrum acquisition), economic forces (data caps), and demand adjustment (WiFi offloading\(^1\) and efficiency improvements\(^2\)).

Notably, there is a growing body of literature pointing to co-location or the proximity of mobile devices to one another as a promising technique for improving the wireless user Quality of Experience (QoE)\(^3,4\). While leveraging co-location is not necessarily new as evidenced by the large bodies of work on opportunistic networking\(^5\) and delay-tolerant networks (DTNs)\(^6\), there are several recent trends that encourage newfound interest in the area. First, newer device-to-device (D2D) technologies such as Bluetooth LE, WiFi Direct, and LTE Direct acknowledge that D2D interactions may not want or need sophisticated user approval / security (e.g. user prompting for pairing). Second,
newer studies conducted on large-scale smartphone populations have demonstrated that co-location tends to be quite prevalent\textsuperscript{7,8,9}.

While knowing that the prevalence of co-location is critical to motivate further research explorations, a key research question is to explore whether or not said prevalence is useful, namely does co-location occur in a manner that can be leveraged in practice? In that context, we focus on one of the more intriguing approaches to leveraging co-location, i.e., that of D2D caching\textsuperscript{4}. In short, D2D caching uses co-location and the respective low-cost of information exchanges of D2D communication to enable mobile devices to share previously downloaded content with one another. Although numerous technical challenges exist (what to exchange, managing dense settings, privacy, object awareness / upstream caching\textsuperscript{10}), foundational questions exist with regards to underlying user behavioral impacts on D2D caching. To start, how fast would shared data propagate through the network? To what extent is there not only propagation but diversity of propagation (shared data from multiple users)? Is storage likely to be a concern? What is the steady-state storage requirements for caching and what impacts do simple tuning parameters such as signal strength or maximum content age play on cache sizing?

It is these questions for D2D caching that our paper seeks to address. We leverage our existing NetSense dataset of nearly two hundred smartphone users with fine-grained co-location, device state, and traffic consumption data to characterize data propagation speeds, cache sizing requirements, and the impacts of co-location quality and data age restrictions. The key contributions of our work are as follows:

- We evaluate real-world data to establish the potential for proximity-based caching on our NetSense smartphone cohort of nearly two hundred users. Our real-world dataset allows us to not only study the rate of propagation and storage requirements but also to capture the nuances of application variations (Facebook, web browser), interface variations (cellular, WiFi), and environmental state variations (phone on / off, signal strength). In contrast to prior works that rely on synthetic models for mobility\textsuperscript{4}, our analysis represents one of the first to study the D2D caching potential in the wild.
- A particularly intriguing definition is the concept of Propagation Volume which captures a representation of the storage costs for exhaustively caching all data as exchanged via all co-location instances with other study participants. In effect, we capture the cumulative cost of caching with volumes couched in actual traffic demand values. The subtle aspect of Propagation Volume is that it allows us to better measure the potential impact of caching by reducing the bias of co-location frequency.
- Our results show that data propagation can occur with the devices having significant portions (90%) of their cached data exchanged. If we consider the actual data volume (i.e. propagation volume), the average drops to just above 70% of the cohort traffic volumes. Moreover, we find that 4 GB or less of storage is nearly sufficient to cache all needed content for an entire day and further find that on average, many smartphones easily have that space to spare and significantly more (roughly 7 GB available per phone in our study).

2. Related Work

The notion of D2D communication is a topic that has received considerable research attention in the community albeit limited deployment in practice. From the perspective of our paper, there are four main categories of related work, namely opportunistic networking, delay-tolerant networks, large-scale smartphone cohort studies, and work specifically focusing on smartphone data optimization.

At its core, current efforts on D2D communication can trace their roots to original research work conducted on opportunistic networking\textsuperscript{11} and delay-tolerant networks\textsuperscript{6}. Opportunistic networking embraces opportunities for nodes to collaboratively take advantage of wireless resources when available as afforded by co-location rather than relying only on a fixed infrastructure. In the typical example, a mobile node will relay the content of another mobile node to grant access to wireless network resources either by virtue of possessing a better path to the wireless network or to extend the range of the wireless network. Conversely, DTNs take opportunities for network access one step further by assuming that all network access will be intermittent and that longer delays are tolerable provided that the data is eventually delivered. In contrast to DTNs, caching in the D2D context has a permanence of state with the goal of content propagation through persistent storage rather than storing and forwarding.
Our own NetSense study is one of several more recent large scale smartphone instrumentation efforts such as PhoneLab, the Nokia Data Challenge, the MIT Reality Mining, and LiveLab. The goal of each of these efforts is to instrument large cohorts of smartphone users (several hundred) for the purposes of studying social interactions with the added bonus of capturing co-location information. In fact, our prior work demonstrated that opportunistic networking has considerable opportunities in the wild and need not necessarily be concerned with device selfishness. Critically, that work focused on only opportunities for co-location relaying or collaboration, not on caching with persistence nor the velocity that data might propagate in such cases.

Newer works have shed important light with regards to optimizing smartphone data consumption. Qian et al. demonstrated that up to 30% of smartphone traffic may be redundant by virtue of poor web browser performance. The AT&T Automatic Redundancy Optimization (ARO) tool offers app designers the ability to specifically measure the amount of redundancy in their data flow for redundancy elimination. More recently, Finamore et al. noted significant opportunities for time-shifting or pre-staging of content by observing web traffic objects from the service provider perspective. Even more recent work by Ji et al. has began to explore the notion of pre-staging / caching from a pure device-to-device perspective. Most importantly though, Ji’s approach works from the perspective of synthetic mobility and demand models while our own work is driven by real-world traces and demand instrumentation of our smartphone cohort.

3. Data Overview - NetSense Study

The data for this work is drawn from our NetSense smartphone study and consists of usage patterns collected by a user-level agent running on the Android smartphone of nearly two hundred campus users. In August of 2011, two hundred incoming freshmen at the University of Notre Dame were each provided a free Android smartphone (Nexus S) and cellular data plan (unlimited data and text) in exchange for complete metadata monitoring rights of all digital interactions (texting, traffic volume, application usage) and phone environmental information (co-location, battery charge, screen state). Critically, we note that actual content was not logged but rather only the metadata itself was recorded for all interactions using the aforementioned user-level agent. For the purposes of this paper, we comment specifically on the analyzed features and their observation frequency with further details left to our overview paper.

The key features include:

- **Co-location**: Each device in the study was configured to be infinitely discoverable with respect to Bluetooth with a condition for study participation being that users must always leave on Bluetooth. Bluetooth co-location traces are recorded in the dataset as a 5-tuple \((deviceID, timestamp, neighborID, neighborMAC, RSSI)\), at the granularity of once every three minutes as the mobile devices invoke Bluetooth neighbor discovery. For the purposes of this paper, the observed Bluetooth co-location instances are filtered to only include devices in the NetSense study.

- **Traffic volume**: Each device in the study records the total traffic (Tx, Rx) across each wireless interface (cellular, WiFi) once per minute. In addition, the data consumption for each application is also logged with respect to downstream (Rx) and upstream consumption (Tx), albeit without interface-specific information.

- **Device state**: Latent storage on each device is recorded as well as the approximate quantities of different data types (e.g. apps, photos, videos) are recorded on the phone once per day. The device state with respect to usage (screen on / off) is recorded on a per-second granularity as triggered by Android system callbacks. Heartbeats as established by the Bluetooth observations ensure the liveness of the user-level agent itself.

4. Evaluation

In order to measure the potential for D2D caching, we developed a discrete event simulator to **play back** the events as recorded by our user-level agent. Rather than focusing on the complexities of fine-grained D2D interactions with consideration for the physical layer, instead we concentrate on higher level properties, namely the extent to which data propagates, the speed and volume of propagation, and the impacts of imposing tuning parameters such as restrictions on D2D signal quality and maximum data age. We begin by first defining several properties of the mobile devices for the purposes of evaluation:
Definition 1: Let MN be the set of mobile nodes and mn_i be the i-th node where mn_i \( \in MN \). |MN| is the total number of mobile nodes in our study.

Definition 2: Let CL_{i,j} be an instance of co-location between mn_i and mn_j at time t, enabling the exchange of any cached data as governed by the D2D caching policy.

Definition 3: For a given mobile node mn_i, let BF_{i} be the block of content downloaded locally on mn_i during the time period starting at t. Hence, for each time period, there are |MN| shareable blocks generated (one at each mobile node). A block can be shared with other nodes via any co-location instance \( CL_{i,j} \) where \( t' \geq t \).

Definition 4: Let \( CP_{i} \) be an indicator that denotes the block generated on mn_j at time period t is present on mn_i. Let the entire set \( CP_{i} \) represent all blocks present at mn_i and downloaded by other nodes during time period t such that \( CP_{i} = \bigcup_{j} CP_{i,j} \). Hence, \( CP_{i} \) represents the set of propagated blocks from other mobile nodes as received either directly or indirectly (through multiple hops).

For the purposes of our evaluation, we set the time period to be equal to one hour thus yielding 24 measurement points in the course of the simulation. The volume (size) of each block is derived from the traffic consumed through the downlink in that particular hour via all wireless interfaces of that mobile device. The user-level agent has only a minimal role as it consumes a limited amount of downlink traffic. Each block is only made available after its respective time for sharing. We note also that a finer-grained period (5 minutes) made little difference with regards to the results of the simulation.

During co-location, a mobile node will exchange its cache with detected neighbors subject to RSSI and maximum age constraints. By default, there are no RSSI filters (RSSI on the Bluetooth link as detected during discovery) nor maximum age restrictions unless otherwise noted. While we acknowledge that placing a cap on the total amount which can be exchanged would be more realistic, our goal is rather to explore the maximum potential for caching, leaving individual data details for future work. In the typical co-location case, a mobile node will try to share its content with mobile neighbors roughly twice per 3-minute Bluetooth monitoring interval (once when it detects the neighbor, once when it is detected as a neighbor). The cache of a given mobile node will then populated both by direct exchanges with neighbors (mn_i receives blocks of mn_j from mn_j) as well as indirect / multi-hop exchanges (mn_i receives blocks of mn_j via mn_k). Both types of exchanges are tracked over the course of the simulation.

4.1. Key Metrics

We propose three metrics to evaluate the content propagation, namely Propagation Ratio, Propagation Latency, and Propagation Volume. The Propagation Ratio (PR) measures to what extent the data blocks generated during a given time period are propagated amongst mobile nodes. \( PR(mn_i, t) \) is a measurement for the blocks created during time period \( t \) which appear later on in the cache of mn_i. A value of 1.0 implies that all data blocks created in t from other mobile nodes were present at the end of the day in the cache of mn_i. The propagation ratio is formally defined as:

\[
PR(mn_i, t) = \frac{|CP_{i}|}{|MN||t|}
\]

whereby the overall system PR can be derived by computing the average propagation ratio across all mobile nodes.

Propagation Latency (PL) as defined below captures the velocity by which a data block propagates across the network. For mobile node mn_i, the propagation latency of a given block created in the time period starting at t is calculated by measuring the arrival time of that block at mn_i versus the original creation time t. The overall system propagation latency is simply the average recorded at each node then further averaged across all mobile nodes. Notably, we break out propagation latency for each individual hour to capture how velocity changes over the day, namely that periods of inactivity or insignificant location changes (e.g. class) can serve to reduce the overall velocity of data propagation.

\[
PL(mn_i, t) = \frac{\sum_{i} arrival_time(CP_{i})-t}{|CP_{i}|}
\]

Propagation Volume (PV) further improves the earlier metrics by taking into account the actual volume consumed rather than whether or not an exchange took place. As noted earlier, one can view this distinction as a key differentiator versus consideration with respect to routing or propagation for DTNs. In effect, the co-location data exchange has pre-staged the content into the cache thereby reducing a later need for data consumption on the expensive primary
The propagation volume captures the actual volume of data that would be shared and is computed as follows:

\[ PV(mn_i, t) = \sum_j volume(CP_{i,j}) \]  

which indicates the total storage required to store, in the cache of \( mn_i \), all of the blocks propagated from other nodes for a particular time period \( t \). More importantly, this figure can be used to derive one additional metric. While the \( PV \) for a given node represents the maximum cache size necessary needed to exhaustively cache data as seen from other mobile nodes without consideration for cache replacement strategy, the ratio of the cumulative \( PV \) to the overall data consumption of all mobile nodes allows one to capture the extent to which data might be cached. At a ratio of 1.0, all data generated in that particular time period would have been shared via co-location while a value of 0.0 represents no sharing. Notably, this does not account for the timeliness of the request (although that can be mimicked by tightening a maximum age) nor necessarily for popularity (e.g., Zipf distribution for data). We feel though that this indicator does shed important light on the potential for co-location caching.

4.2. Results

4.2.1. Propagation Ratio

We begin with our evaluation by first exploring the propagation ratio. Fig.1 plots the propagation ratio constrained by signal strength (RSSI). We apply different levels of RSSI to investigate the potential impact of the channel quality on content propagation. The value of -80dBm is commonly used to indicate feasible communication through wireless while -65dBm indicates “close” distance between devices\(^{17}\). When a specific RSSI filter is applied as the constraint, only co-locations with greater or equal RSSI values are used for content exchange. In the case when the RSSI is insufficient, co-location cache exchanges are not allowed to occur.

In the graph, multiple weekdays are selected to evaluate the propagation ratio. We choose weekdays since class days normally imply additional co-location opportunities due to campus mobility as opposed to non-class days. Equation (1) is applied to 24 measurement points for each day and the average case is illustrated in Fig. 1. The results demonstrate that in general, for a -80dBm RSSI restriction, the content downloaded before noon has a higher propagation ratio (\( \geq 0.9 \)) compared to the content downloaded in the evening (\( \leq 0.65 \)). One important reason lies in that content generated in earlier time of the day has a longer period to propagate and thus could potentially leverage more co-location opportunities. The ratio decreases radically when RSSI is constrained to -65dBm, indicating that in our dataset co-location samples with “close” distances have fewer records than with feasible RSSI values (-80dBm). Most notably, reasonable rates of content exchanges can still occur closer between -80dBm and -65dBm but -65dBm affords higher bandwidth and / or lower power communication versus the relatively weaker RSSI values.

4.2.2. Propagation Latency

The average case of propagation latency for the same weekdays is plotted in Fig. 2. When the RSSI filter is set to -80dBm, the blocks created from traffic at 12AM (hour 0) take roughly 11 hours to propagate, which is expected since
content exchanges rarely occur during sleeping hours due to the rareness of co-location. In contrast, blocks generated between 8AM and 3PM are propagated in about 3 to 4 hours due to relatively frequent D2D interactions. When one corrects for the duration of sleep (estimating 7 hours), the two instances of propagation are largely congruent.

At first glance it might be surprising that data generated after 5PM are propagated relatively faster (2 hours). However, as indicated by Fig. 1, the propagation ratios for content generated in the evening are less than 0.65, thus yielding fewer values contributing to the average latency as a hard stop is applied at the end of the day. If one stretches the evaluation time to 48 hours rather than the 24-hour period of the above graphs, similar diurnal patterns are observed.

Table 1. Storage Requirement per Device (MB)

<table>
<thead>
<tr>
<th></th>
<th>Day 1</th>
<th>Day 2</th>
<th>Day 3</th>
<th>Day 4</th>
<th>Day 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>3461.6 (0.73)</td>
<td>3056.1 (0.72)</td>
<td>3890.2 (0.73)</td>
<td>2920.8 (0.77)</td>
<td>2494.2 (0.74)</td>
</tr>
<tr>
<td>Median</td>
<td>3575.4 (0.76)</td>
<td>3278.5 (0.77)</td>
<td>4042.5 (0.76)</td>
<td>3037.5 (0.81)</td>
<td>2638.8 (0.78)</td>
</tr>
<tr>
<td>Maximum</td>
<td>4068.1 (0.86)</td>
<td>3469.7 (0.81)</td>
<td>4570.5 (0.86)</td>
<td>3356.8 (0.89)</td>
<td>2988.2 (0.89)</td>
</tr>
</tbody>
</table>

4.2.3. Propagation Volume

Fig. 3 and Fig. 4 respectively demonstrate the two volume metrics proposed earlier, using RSSI as a filter. Fig. 3 shows that content downloaded around 1PM yields the highest propagation volume (over 350 MB with RSSI -80dBm). This can be partially explained by the fact that mobile users in campus typically consume more network traffic during the lunch break as compared to the cases for morning and afternoon when the students are attending classes. The other reason could be the high propagation ratio suggested in Fig. 1, i.e., around 0.9 for noon hours. Interpreted another way, from the standpoint of a mobile node we consider the average storage requirement for caching blocks that are generated at 1PM and propagated from other nodes as 350 MB.

Table 1 summarizes statistics of the storage requirement for an individual mobile node to cache blocks generated and propagated through the whole day. Values in the parentheses represent the ratio of the storage cost to the overall daily traffic. The maximum requirement is 4570.5 MB which is feasible for today’s smart devices. Notice that Table 1 indicates the upper-bound requirements since in our simulation mobile nodes share all content via co-location, the storage cost is expected to decrease with more sophisticated exchange protocols.

From Fig. 4 we observe a similar distribution of volume ratio, compared with the propagation ratio plotted in Fig. 1. We note that for content generated between 12PM and 1PM, the propagation volume ratio remains around 0.9, indicating considerable potential for eliminating redundant network traffic during peak hours (see Fig. 8) via content exchanges. Compared to the case of noon hours, even though content downloaded in the morning has slightly higher volume ratios, the raw propagation volumes are lower as the potential benefits gained from D2D sharing for the morning content is actually less.
4.2.4. Impacts of Content Age

In addition to the impact of RSSI, we also explore the performance impacted by different content age constraints. Age indicates the potential utility of shared content, since the faster a piece of content is propagated the more likely it arrives at nodes requiring the same content in a timely manner such that retransmission from the networking can be avoided. Specifically, when an age constraint is set, only content with equal or smaller age than the selected value is allowed to be exchanged.

We use three levels of age restrictions, namely 2, 4 and 8 hours, and plot previously-defined performance metrics (omitting PL) respectively in Fig. 5 trough Fig. 7. Fig. 5 shows that even with the most relaxed 8-hour constraint, midnight content yields a propagation ratio of only 0.1, which is in accordance with the observation from Fig. 1 that blocks created at 12AM generally have a propagation latency greater than 10 hours. For content generated at peak traffic hour (12PM-1PM), the tightest 2-hour constraint still yields a nearly 0.6 propagation ratio, suggesting acceptable overall performance.

For propagation volume as shown in Fig. 6, taking the content created during noon hours (12PM, 1PM) as an example, the values are 250-350MB, 190-200MB and 80-150MB when 8, 4 and 2-hour age constraints are applied respectively. This observation indicates that most of the volume propagated to mobile nodes are contributed by content with age ≥ 4 hours. Again, propagation ratio and volume ratio have similar distributions when using the same age constraint, indicating a strong correlation between the portion and the raw volume of propagated content.

In general, our results show that it takes roughly 4 hours for the content generated between 8AM and 3PM to disseminate. In addition, the propagation performance significantly decreases when using 2-hour as the age constraint. Further investigation on potential reasons for relatively high latency and performance drop will be conducted in future work, and is skipped in this paper due to space constraints.
5. Conclusions

In this paper, we evaluated the performance of content propagation via device-to-device communication by using realistic co-location and network traffic data. The initial results, particularly the propagation ratios in terms of both block count and raw volume, demonstrate that there exists considerable potential for D2D content caching through relatively good channel quality (-80dBm). Therefore, we posit this particular domain is worth more attention from the research community. Although part of the performance measurements suggest long propagation latency, we believe the related results were affected by our exchange strategy, i.e., accounting all blocks from co-location neighbors.

Looking beyond our work contained in this paper, several open areas remain for further exploration. First, the social network relationships among mobile users could be investigated for inferring content of potential common interests. Second, consideration of certain caching policies might lead to significantly different results, while in this paper we adopted a simplified approach by exchanging all shareable content. Last but not the least, applying similar evaluation on a larger set of user pool could also impact the propagation performance due to a different density of Bluetooth co-location, especially when considering the two-hundred user pool is only 2.5% of the overall population of undergraduate students in campus.

Acknowledgments

This work was funded in part by the National Science Foundation through grant IIS-0968529. We would also like to thank our collaborators, Prof. Christian Poellabauer, Prof. David Hachen, and Prof. Omar Lizardo. Special thanks to Sprint who provided the ability to offer the phones and service plans for the study.

References