Measurement and analysis of the spatial locality of wireless information and mobility patterns in a campus

Mark R. Lindsey, Maria Papadopouli, Francisco Chinchilla, and Abhishek Singh
{lindsey,maria,fchinchi,asingh}@cs.unc.edu

March 12, 2003

Abstract

An environment is characterized by *spatial locality of queries and information* when it is likely that users in close geographic proximity query for similar data. Information exhibits spatial locality when it is coupled to a real-world place. For example, play reviews are most relevant in a theater; and users in a Dental school may be particularly interested in web sites on related subjects. The prevalence of such spatial locality is related directly to the feasibility of deploying location-dependent services. Intuition suggests that a high degree of spatial locality of information exists because people often gather to exchange information. As such, we expect that appropriate location-dependent services may harness the spatial locality effectively.

The web is not primarily a location-dependent service, but it provides a ready testbed to study the prevalence of spatial locality and mobile users. This paper results from a three-week study of spatial locality phenomena among mobile web users on a major university campus using the 802.11 [11] wireless infrastructure. We show that users are often near other users with similar interests. In addition, we categorize the URLs and present a classification of the wireless information as a function of the location from which it was accessed. We also model the associations of wireless users to access points. Finally, we discuss the implications on the feasibility of location-dependent services and potential improvements of the wireless access using caching mechanisms.

1 Introduction

Wireless devices are becoming smaller, more user friendly, and more pervasive. They are not only carried by people, but are integrated into stationary objects. These devices can be part of data-centric, mobile ad hoc (i.e., constructed without an infrastructure) and sensor networks; they collect, measure, process, query, and relay information. Mobile users access news, traffic or weather reports, maps, video files, games, and information related to events in close geographic proximity. At the same time, handheld devices have limited energy and memory, and relatively low communication bandwidth. In addition, users may need to discover information while mobile in a dynamic environment. In that case, prefetching the data prior to the disconnection of the device would not be possible. Based on their dependency on an infrastructure, we can classify the mobile data access into three categories, namely via an infrastructure of base stations (e.g., via an IEEE 802.11 [11] deployment of APs), information servers (e.g., infostations [10]), and peer-to-peer (e.g., 7DS [14]).

The peer-to-peer approach introduces a new paradigm of information sharing and cooperation among mobile devices not necessarily connected to the Internet. The devices collaborate in a self-organizing manner without the need of an infrastructure. 7DS is a system that uses this mechanism. The wireless data access to the Internet via base stations is characterized by frequent disconnection and low bit rates. When a wireless infrastructure to the Internet is available, it operates as a proxy and forwards user queries to the Internet. However, when a host experiences loss of connectivity, 7DS attempts to acquire the data from peers within its wireless coverage by broadcasting a query. Due to the highly dynamic environment, the system does not try to establish permanent caching or service-discovery mechanisms. Instead, it exploits the transient aspect of information dissemination, and the spatial locality of queries and information. An environment is characterized by spatial locality of queries and information when it is likely users in close geographic proximity to query for similar data.

Papadopouli *et al.* showed via simulation that in settings with high spatial locality of information, these peer-to-peer systems can enhance the information access by reducing the average delay to receive the data and increasing the probability of discovering the data. This study seeks to examine the feasibility of such

systems; *if a client could retrieve objects from others nearby, how often would that be helpful?* An ideal approach to this question requires that we compare the information requirements of users to the information availability from those nearby; we approximate this by correlating mobile users' requests for web objects with their location, and with the request history of other users nearby at the time of the request. Because there are many web sites on any common topic, we also compare the categories of information which are requested. We further examine the overall popularity distributions of objects requested by mobile users. Finally, we examine patterns of mobility among wireless users to provide intuition on the movement behavior. We study these issues by recording the HTTP requests and movement of mobile users on a university campus over a three-week period.

We differ from previous similar efforts by focusing on the activity of individual clients rather than on the entire population of mobile clients. Unlike previous studies of wireless networks, we explore the effects of spatial locality of information by studying the information requirements of mobile clients. And instead of using simulation to model user mobility and data dissemination patterns, we examine the behavior of real clients moving in a real wireless network.

We show that there is opportunity for improving adhoc peer-to-peer systems and wireless data prefetching systems by focusing resources on maximizing data availability over short time periods, and minimizing the extended storage of data when it has not been recently requested. We find that there is a significant correlation in the general interests of users who are near one another: among requests for objects known to be in a specific category, 22% of requests by mobile users are for information in the same category that another nearby user has requested within the past hour. Also, most mobile users (55%) are relatively stationary, visiting only a single access point each day.

Section 2 of this paper describes previous related research; Section 3 explains our techniques for acquiring the data for this study; section 4 describes our study of spatial locality and web objects; section 5 gives findings on popularity distributions for objects and categories; finally, section 6 provides insight about the movement of users on the campus.

2 Related work

Previous projects, such as 7DS and VIA [14, 7], motivate the idea of exploiting user mobility for location-dependent data propagation. They built a foundation of architectural and theoretical feasibility of such systems, and we further examine the feasibility in the context of such object distribution system in the context of web objects and mobile wireless users.

Collaborative caching among conventional non-mobile clients has been analyzed by Duska *et al.* [8] They find the benefit of such caching to be limited by the diversity of clients' requests and the non-cacheability of many objects. Intuition leads some to believe that the overlap between clients' requests may be greater among mobile users, and we examine this hypothesis. Further, we consider only an *ideal* cache, in which objects have an arbitrary useful lifetime, as we envision a day when web objects – like printed matter – are informative indefinitely.

There have been other studies of client mobility and access patterns. Bhattacharya and Sajal [5] performed a similar study of client mobility patterns using a PCS network, and propose a prediction mechanism based on Markov states. Kotz and Essien [12] characterized Dartmouth's wireless network, examining global traffic and access-point (AP) utilization. Balachandran *et al.* performed similar measurements in a three-day conference setting, also focusing on the offered network load, and global AP utilization [4]. We build on the work in these papers by considering traffic, but we direct our attention to web traffic. Instead of looking at global patterns of mobility, we focus efforts on modeling an individual user. And unlike any of these studies, we examine the spatial locality of information.

3 Data acquisition

Three sets of data were used to perform this study: traces of mobile client's web requests, categorization data for well-known web sites, and logs of 802.11 MAC events generated by wireless access points in the campus.

3.1 Definitions

The campus is populated by people who have devices which communicate with the campus wireless network; each such device is called a *client*. The campus has many *Access Points*, or *APs*, each of which is a non-moving bridge between the conventional campus network and the wireless network. Each AP has a coverage area determined by radio propagation properties around the AP; a client communicates via the network by establishing a *session* with an AP; we say synonymously that such a client *visits* the AP. The session ends when the client notifies the AP of its departure, or the AP detects that the client is inactive.

3.2 Campus wireless network

The University of North Carolina at Chapel Hill began deployment of an IEEE 802.11 [11] network in 1999. Wireless access is available in many residence halls, academic buildings, the medical school, and in some off-campus administrative buildings. The campus uses primarily Cisco *Aironet 350* 802.11 access points [1], although some areas on campus are serviced by older APs from other manufacturers.

We observed 6,709 distinct mobile users during the period January 24 through March 6, 2003. Of these, 2,494 were observed to make HTTP requests during the tracing period February 6-24, 2003.

3.3 HTTP traces

The bulk of the campus wireless network has a single aggregation point which connects to a gateway router. This router provides connectivity between the wireless network and the wired links – including all of the campus computing infrastructure, and the Internet. We connected to a *monitor port* on the gateway router, letting us monitor all of the traffic that passed between the wireless network and conventional wired networks.

This tap link was connected to a FreeBSD monitoring system. We used the tracing tool *tcpdump* to collect all TCP packets which have payloads that begin with the ASCII string "GET" followed by a space. The full frame was collected as a potential HTTP [9] request. We did not restrict our collection only the standard HTTP port so that we could record HTTP requests sent to servers on non-standard ports, which includes many common peer-to-peer file-sharing applications.

The packet trace was then processed to extract the HTTP GET requests contained therein. From each packet, we keep these items:

- The time of the packet's receipt (with one-second resolution),
- The hostname specified in the request's *Host* header [9],
- The Request-URI, (e.g., `/mobicom/2003/`)
- The hardware MAC address of the mobile 802.11 client.

If all of these items were not available in a packet, then we did not include the recorded packet in our recorded requests. 6,319,272 requests were traced and included in analysis.

By recording the traffic before it had passed through an IP router, we were able to capture the original MAC header as generated by the 802.11 clients for transmission to the gateway router.

To avoid violating our users' expectations of privacy, we do not store the hostname, path, and client MAC address directly. See Section 3.6 for details on the techniques used to ensure privacy is maintained.

3.4 Category information

We were interested in measuring instances of correlation between clients or places and queries for information. Information on a particular topic may appear on many web sites and at many URLs, so if we measure only cases where clients request exactly the same web objects, then we would miss many cases where they are requesting web objects that are substantially similar. For example, two clients near one another may consistently access sites which have current news, but they may access different news sites. We sought to measure this correlation.

Further, we expect that users may be flexible in the source of their information in the absence of access to the ideal source. Consider by comparison a person in a coffee shop: he may prefer the *Wall-Street Journal*, but may settle for the *New York Times* because a copy is available. The same user with a mobile device may be flexible in the source of his information. Categorization lets us see past the precise object identified by a URL to identify the nature of information which is requested.

To do so, we attempted to categorize each request which was observed. We used the complete *Open Directory Project* (ODP) database [2], a human-edited index of web sites and categories, as of February 1, 2003. The ODP organizes sites into a hierarchy of categories; the top-level categories include "News", "Society", "Home", "Arts", "Kids and Teens", *etc.* Each of these top-level categories contains both web sites and other sub-categories; e.g., the "News" category contains sub-categories "Media", "Weather" and also popular news sites.

The ODP database that we used indexed 3,181,773 different sites. The ODP categories can reveal much about the sites that were accessed; to protect the privacy of our campus users, we did we chose not to include the full detail of the ODP in our data. Thus, we chose to limit the depth of category detail to three levels of the ODP; for example, this includes *News: Weather: Air Quality*. The names of longer categories were truncated to three levels; for example, all sites in the ODP category *Regional: North America: United States: Government: Executive Branch* were considered to be in the category *Regional: North America: United States* instead. This technique yields 6,851 distinct three-deep categories.

The ODP only lists the URL for the top-level page for a site; most sites have numerous objects referenced from that page. For each recorded request, we truncated the

virgule-separated path elements from the end of the request until we found a match, or had removed all of the path elements without finding a match.

For example, if a request for "`http://www.acm.org/sigmobile/confs/`" was recorded, then we would attempt to find an exact match for this URL in the ODP. If one could not be found, then we would look for a match for "`http://www.acm.org/sigmobile/`". This entry is present in the ODP under the category *Computers: Computer Science: Organizations: ACM: Special Interest Groups*. However, since we are recorded only three levels of depth from the ODP, we would record that this request was matched to the category *Computers: Computer Science: Organizations*. 29% of all recorded requests were matched to one or more categories using this procedure. Many sites appeared in several categories; if the site appeared in fewer than eight categories, then we included it in each of those. This left a few sites which appeared in a suspiciously large number of semantically diverse categories; in these cases, we considered the URL not to be categorized at all.

3.5 Access-point logs

The campus primarily uses Cisco Aironet 350 802.11 access points (APs) to provide the wireless network service [1]. These APs can generate log messages for 802.11 MAC level events, which indicate when a user enters the range of the AP (i.e., *associates* with it) or leaves its range (i.e., *disassociates* from it). The majority of APs on campus were configured to send this data via *syslog* to a server in our department. The messages sent by the APs are detailed by Kotz and Essien [12].

We recorded 731,866 useful MAC events, and recorded mobility among 5,479 802.11 clients. 1,460 users were active among 193 APs on an average day during the trace period.

3.6 Privacy assurances

To avoid disclosure of the identity of individual users, and of the sites that a user is visiting, we store and use SHA1 [13] hashes of the client's MAC address, the request hostname, and the requested path. The MAC address uniquely identifies an 802.11 network device; we assume it to be coupled to a specific computer. Two requests are considered to be from the same client if they were generated by clients which have the same hashed MAC address, and two requests are considered to be for the same URL if they have the same hashed hostname and the same hashed requested path.

4 Spatial locality and web objects

Four principle methods of analysis are described here:

HTTP request rate over time

Same-client repeated request is a GET request for a URL which was made by a client which had requested the same URL at some time in the past.

Same-AP repeated request is a request for a URL made by a client within an AP such that the same URL had been requested in the same AP at some time in the past, possibly by a different client.

AP-coresident-client repeated request is a request for a URL made by a client within an AP's area, such that another client that is present in the AP's area at the time of the request has requested the URL at some time in the past. The other client may have been elsewhere when it requested the page first.

AP-coresident-client repeated category is a request for an object which is known to be in some category such that another user that is present in the AP's area at the time of the request has requested an object of the same category at some time in the past.

Repeated requests and correlations among other users are not represented for the first four days of the trace, as these may include significant startup correlation phenomenon akin to compulsory cache misses.

4.1 HTTP request rate

Wireless users' HTTP requests were recorded from 2:31pm, Thursday, February 6, 2003 through 2:51pm, Monday, February 24. During this interval, 6,319,272 HTTP GET requests were recorded successfully for use in this analysis.

Figure 1 confirms similar traces of network traffic, displaying daily patterns of activity which tend to peak toward midnight. Web use on the wireless network decreases considerably on the weekends (February 8-9, 15-16, and 22-23).

4.2 Same-client repeated requests

Same-Client Repeated Requests are those cases where a single client is observed to request an object which it has requested in the past. The cause could be any of these:

Figure 1: Recorded HTTP Requests.

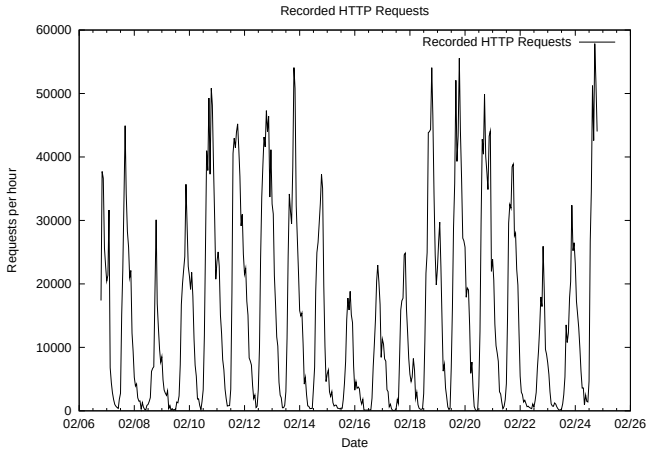
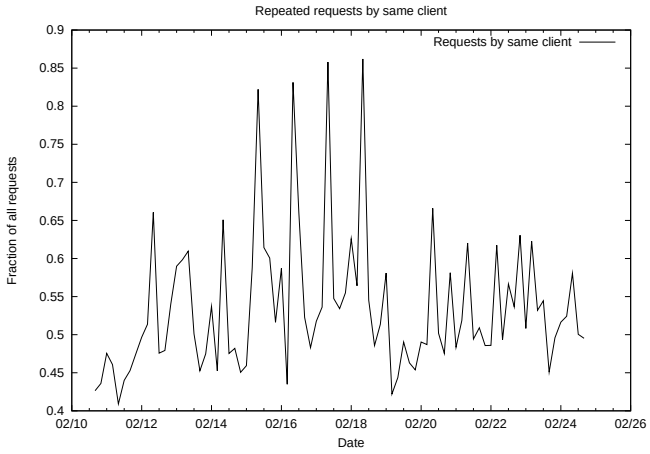


Figure 2: Same-client repeated requests.



Subsequent request. A user intentionally requests an object which they have requested in the past, but which could not be satisfied by the browser cache. Such a request would represent genuine ongoing interest by some user.

Automatic reloads. Many popular pages (e.g., headline-news sites) cause the browser to re-load the page periodically [1]. While the page is displayed, the browser will periodically re-request it. Such requests could also be considered indicative of continued interest by the user.

Packet retransmissions. The first packet containing the request was not known by the client to have reached its destination, so TCP specifies that the client will retransmit the packet. We would record both requests as distinct requests. We expect that such retransmissions are rare: The primary source of retransmissions is packets dropped in the queues of congested routers. The GET request typically occurs in the first packet of the TCP session; at that stage in TCP, the client only has a single packet in the network, so it is minimally subject to drops due to congestion.

This study is subject to the effects of browser caching; if the requested object is in the browser's cache, then no HTTP request will be generated. Some, but not all, browsers follow HTTP's specification for determining the freshness of a cached object.

This measure does not account for the location of the client. Because we did not have complete information about all APs in the wireless area, we could not always be certain of the location of a client when its requests were observed.

Figure 2 shows the variation of same-client repeated requests over time as a fraction of all requests observed in the hour. Figures 3 and 4 show the frequency of repeated requests. These graphs can be interpreted as answering: what fraction of all requests were requests for an object which had been requested by the same client h hours in the past, where h is the horizontal axis? For example, over 35% of all requests were for objects which had been requested by the same user within the past hour; this suggests that as many as 35% of all requests would be unnecessary if every object on the web had a cache lifetime of at least an hour, assuming that all browsers observe the HTTP standard for caching.

Figure 3 introduces an unsurprising trend present in all of the long-range frequency-domain visualizations shown: 24-hour cycles.

4.3 Same-AP repeated requests

When an object is requested multiple times within the same AP's area, then those are called same-AP repeated requests. This measure does not account for the client which makes the request; i.e., the repetition can occur because of a single client's activity within a single AP's area, or because of several clients requesting the same object.

Figure 5 shows the variation in such same-AP repeated requests with respect to time. Figures 6 and 7 show the fraction of requests for objects which had been requested within the past h hours, where h is the horizontal axis. The *overall* information presented describes the overall probability of all requests where the AP of the requesting client was known. The APs of several

[1] This is accomplished by the META HTML tag included in the HEAD Section of a specified with HTTP-EQUIV="Refresh".

Figure 3: Same-client repeated requests over 12 hours; each bin represents four hours.

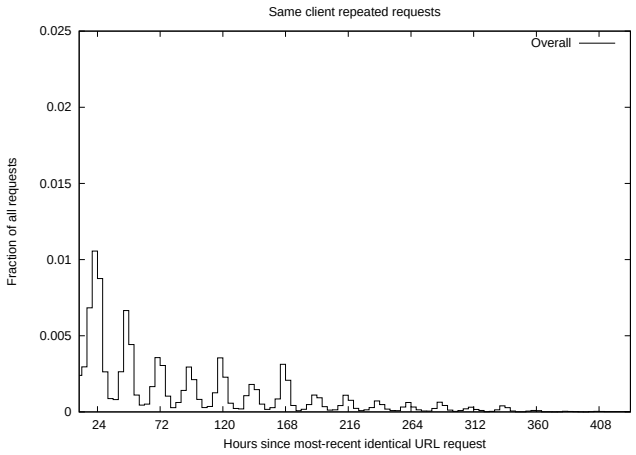


Figure 4: Same-client repeated requests.

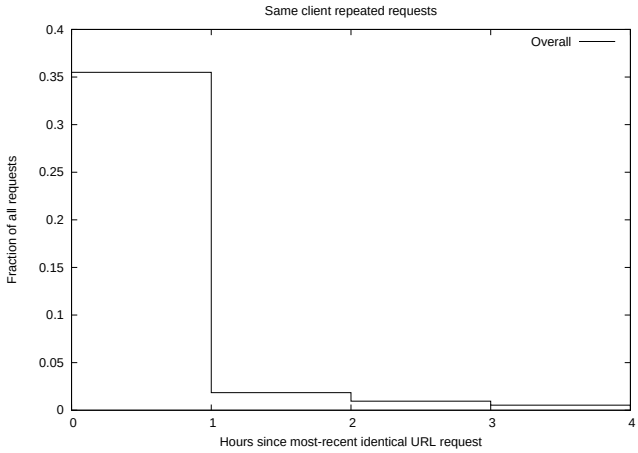


Figure 5: Repeated requests within a single AP.

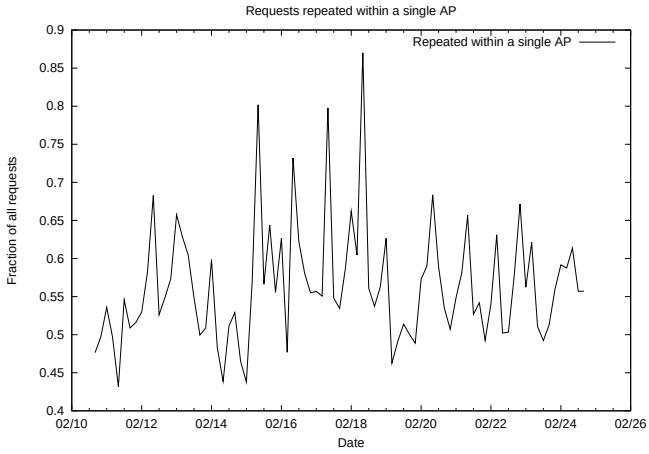
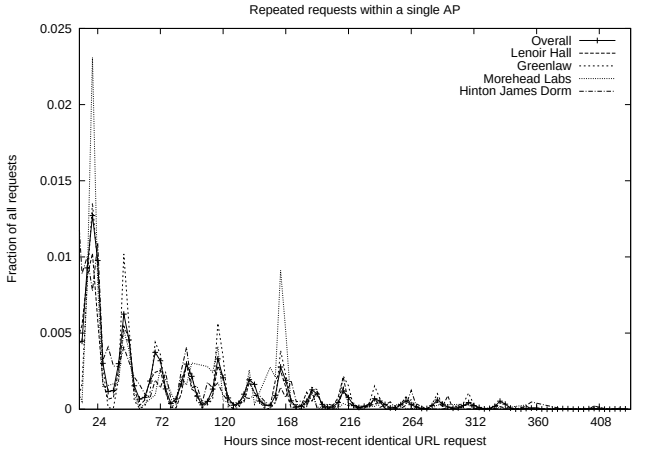


Figure 6: AP-client repeated requests over 12 hours.



other buildings were sampled, and their representative descriptions are shown. A significant number of observed requests were available for each location shown.

4.4 AP-coresident-client correlations

At the heart of measuring spatial locality effects among mobile web users is this: how often are users who are interested in the same things near one another? We answer this question in two ways: by examining object correlations among the objects requested, and correlations among the categories for requested objects.

In general, such correlations occur when a client in an AP's area requests an object which is *related to an object* which has been requested at some time in the past by another client in the same AP's area at the time that the new request is made. Requests are considered to be related when they are for the same object (i.e., the same

Figure 7: AP-client repeated requests.

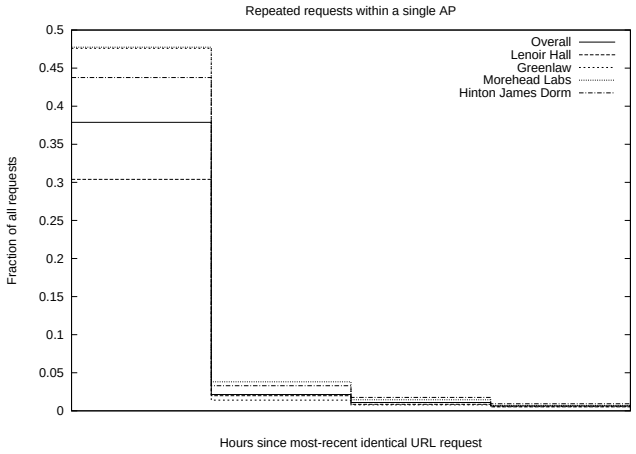
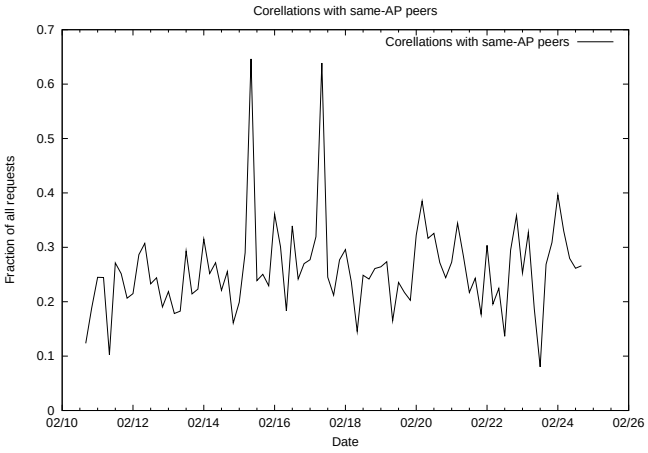


Figure 8: AP-coresident-client correlated requests.



URL), or for the same category. These two methods of comparison are describe below.

Figure 9: AP-coresident-client correlated requests over 12 hours.

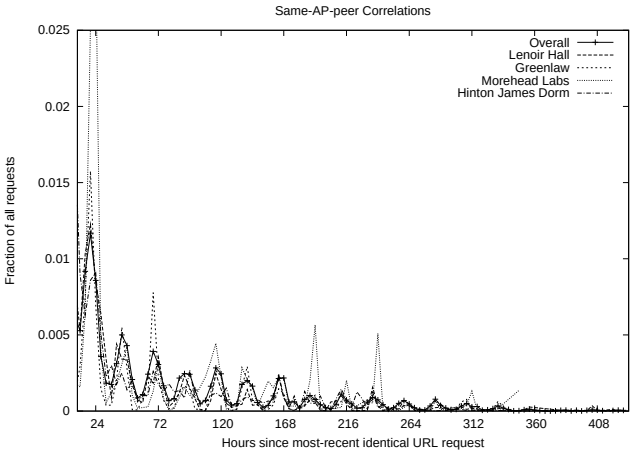
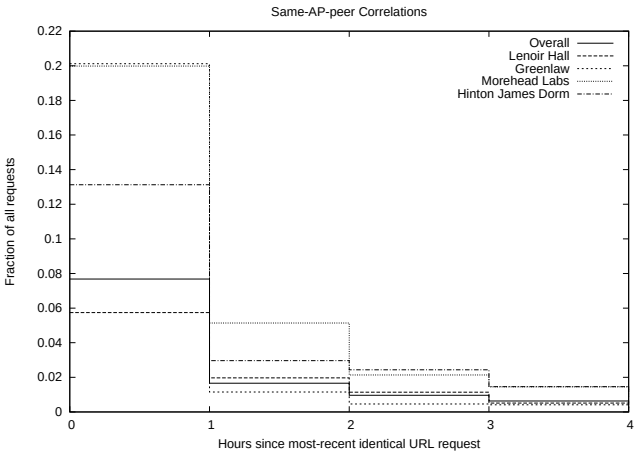


Figure 10: AP-coresident-client correlated requests.



4.4.1 Same-URL requests

An AP-coresident-client same-URL repeated request is said to occur when a client in an AP's area requests an object which has been requested at some time in the past by another user who is in the same AP's area at the time that the new request is made. Note that this other user, who requested the object in the past, may have requested the object while at a different location. This indicates that two different users have requested the same object, but were near one another at time of the second request. Figure 8 shows the relative occurrences of this over the course of the trace. Figures 9 and 10 present the proportion of such requests with respect to time since the earlier requester requested the object. If a client requests an object while in the AP with two or more other clients which have also requested the object in the past, then the latest of the earlier requests is used to establish the interval between requests.

We observe that over 7% of all requests are for objects which have been requested by a nearby user within the last hour. Furthermore, this proportion varies widely; at some locations on the campus, over 20% of all requests were for such objects.

4.4.2 Related-category requests

Our goal was to detect instances that users had related information needs. For most topics, there are several web sites that contain information about the topic, and each web site may include several pages, and each page may contain several objects. Thus, even when nearby

Figure 11: AP-coresident-client correlated category interests.

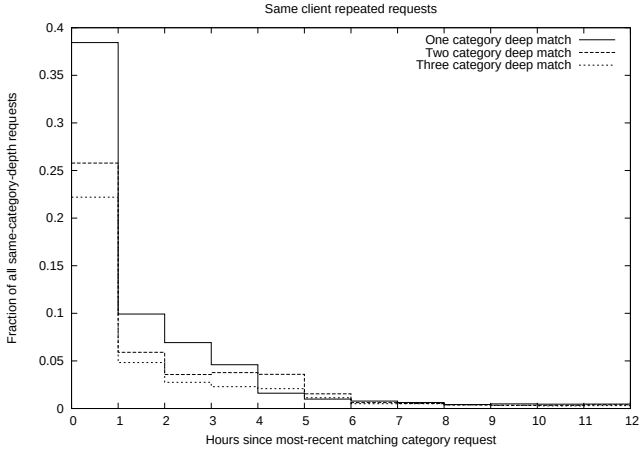
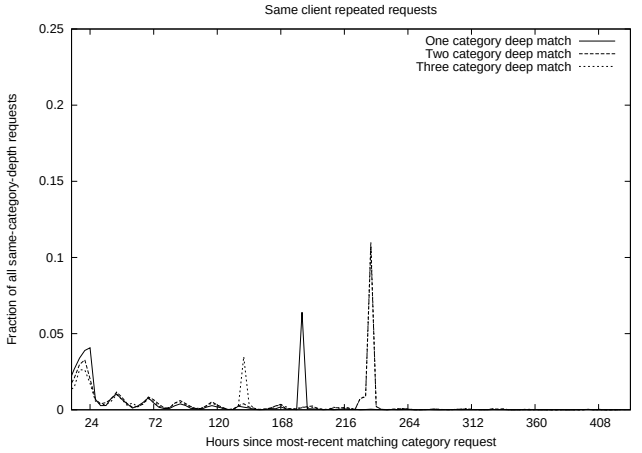


Figure 12: AP-coresident-client correlated category interests.



clients are interested in the same things, instances of overlapping requests for objects may be relatively rare.

As described in Section 3.4, we attempted to associate each request to one or more categories; approximately 28% of all requests could be matched to at least one category. Then we analyzed to see how often AP-coresident users were interested in the same categories. Figures 11 and 12 reveal that 22% of all requests for objects known to be in a category were made while the client was in the same AP with another client which had requested an object in that same category three-deep within the last hour.

Rank	Requests	Category
1	362,690	*Reference: Education: Colleges and Universities*
2	196,473	*Regional: North America: United States*
3	86,700	*Computers: Internet: Searching*
4	67,694	*Business: Major Companies: Publicly Traded*
5	63,239	*Arts: Television: Networks*
6	53,484	*Business: Arts and Entertainment: Media Conglomerates*
7	48,900	*Sports: Resources: News and Media*
8	28,106	*Regional: Europe: United Kingdom*
9	25,690	*Health: Dentistry: Education*
10	23,151	*News*
11	22,386	*News: Newspapers: Regional*
12	22,246	*Recreation: Travel: Consolidators*
13	19,611	*Computers: Internet: On the Web*
14	17,641	*Arts: Television: News*
15	16,197	*Reference: Libraries: Library and Information Science*
16	15,473	*Computers: Software: Internet*
17	15,051	*Reference: Dictionaries: World Languages*
18	11,945	*Computers: Internet: Access Providers*
19	11,921	*Arts: Music: Styles*
20	11,762	*Shopping: Varied Merchandise: Major Retailers*
21	11,542	*News: Breaking News*
22	11,431	*Computers: Software: Operating Systems*
23	10,443	*Society: Law: Legal Information*
24	10,314	*Computers: Internet: Web Design and Development*
25	10,119	*Science: Social Sciences: Communication*

Figure 13: Top Categories

5 Object and category popularity

This section describes the popularity distributions of requests for objects, and of requests for categories. Figure 13 shows the top 25 categories. Note the sharp decline in the number of requests with rank. The university's site is included in the *Reference: Education: Colleges and Universities* category, which constitutes the most-popular category. The category *Regional: North America: United States* encloses a large number of general-interest sites; after that there is a slower decrease in the number of requests for the categories.

Sites for entertainment and news and the related categories constitute a substantial proportion of the most common categorized requests. Specialized categories like *Reference: Libraries: Library and Information Science* and *Health: Dentistry: Education* appear high in ranking, likely due to the nature of the studied population.

5.1 Ranking distribution

The Zipf distribution describes of occurrence of some event P as a function of the rank i when the rank is determined by the frequency of occurrence as a power-law function $P_i \propto i^{-\alpha}$ with the exponent α close to unity; P_i is the probability of the i^{th}-ranked event occurring [6]. The URL rank distribution Figure 14 has the parameter α as 0.85 (denoted as 'a' in figures) [6]. The rank distribution of the servers, shown in Figure 15, has a Zipf-like distribution with α as 1.0. Many of the highly-ranked URLs belong to the same highly-ranked web server.

While the URL graph catches up with the Zipf distribution the web server graph then dips. This indicates that these are the highly-ranked servers which host a few highly-ranked URLs to them. The web server graph

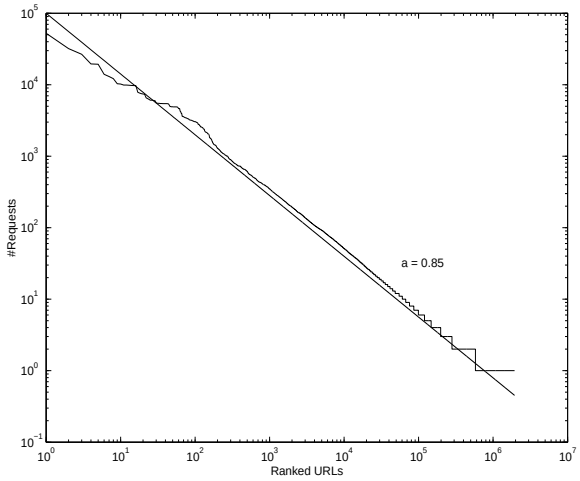


Figure 14: URL Distribution; a = α.

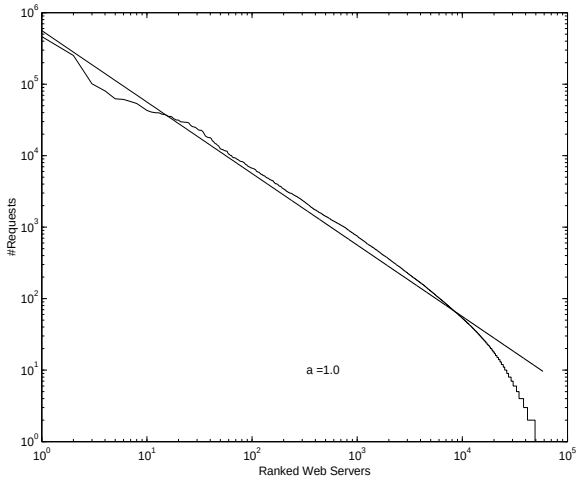


Figure 15: Web Server Distribution; a = α.

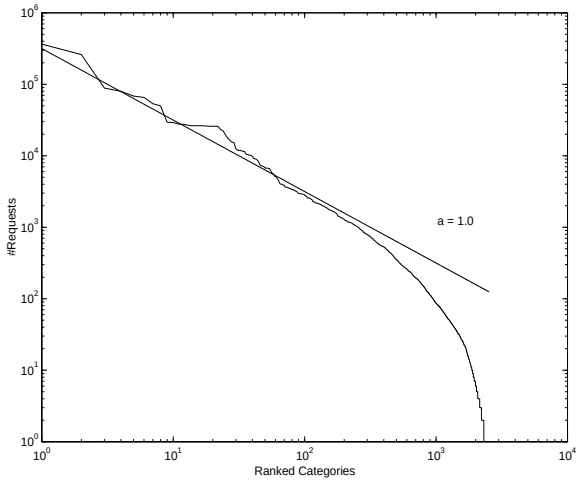


Figure 16: Category Distribution; a = α.

Figure 17: Descriptive statistics for the average daily number of APs and clients.

	APs	Clients
μ	193.3	1310.48
σ	24.63	491.91
$C.I._{95\%}$	193.3 ± 10.53	1310.48 ± 210.39

and the URL graph now follow nearly the same pattern when the web server graph falls of smoothly deviating from the straight Zipf line. There can be varied reasons for this fall. One of the factors can be that, for the low-ranked web servers, the number of requests falls off much more rapidly. This can be attributed to the fact that the higher-ranked servers had certain URLs that were low ranked. We observed 1,941,520 distinct URLs in 2,530 different categories, residing on 58,486 distinct web servers.

The category rank distribution shown in Figure 16 matches the web server graph more. This is because grouping together the URLs is a kind of discrete classification. The category graph is a bit jumpy where it sways and catches up again with the Zipf with $\alpha = 1.0$. One of the reasons for the jumps is that that a URL can be categorized into more than one category, so some categories can show correlated increase. Then there is the same smooth drop as with the web server Figure 15 but this time it is more pronounced. The large number of un-categorized requests contributes to the fall as one of the reasons that these URLs were not classified is that the categories do not cover the entire URL space and there are some URLs which do not fit into the defined categories.

6 Mobility patterns

This section discusses the patterns of mobility among users on the campus. We show here which AP areas have the most distinct clients, how many clients visit many AP areas, and how many clients have long sessions.

Figure 17 presents the mean, standard deviation, and the 95% confidence interval for the average number of APs and clients observed to be active each day our trace.

6.1 Access-point visits

In order to investigate which APs are visited by the most clients, we measure the number of different clients that start sessions with an AP during each hour. Figure 18 illustrates the number of distinct clients which start sessions during each hour; on average, 2.39 distinct clients start sessions with each AP in the noon

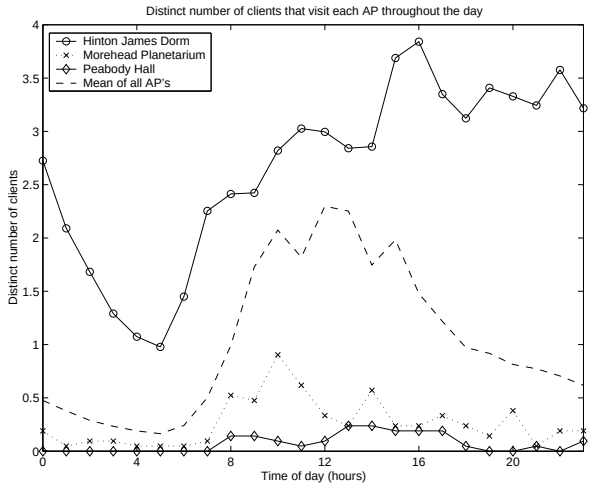


Figure 18: Number of new sessions at selected locations versus time of day.

Figure 19: Example of a client's mobility pattern. *exits* indicates the client's departure from the network.

Time	10:58	11:25	11:57	12:20	17:20	17:25
AP	1	2	1	*exits*	3	*exits*

hour. The APs located in Hinton James dorm, Morehead Planetarium, and Peabody Hall (where the School of Education is located) are indicative of the major trends that appear in our results. Results for these APs and the mean of all the APs are shown in Figure 18.

6.2 Mobility of clients

We also study the number of different clients who visit different APs during time intervals of different length. We consider (separately) non-overlapping time intervals of 20 minutes, 1 hour, 3 hours, 6 hours, 12 hours and 24 hours. We count the number of distinct APs that a client visits during the chosen time interval. If there are multiple time intervals of the same length within a day, the number of clients is averaged for those time intervals.

As an illustrative example, consider a time interval of 12 hours starting at midnight, and the mobility pattern shown in Figure 19. This client is recorded as having visited two distinct APs in the first 12 hour interval: AP 1 and AP 2. Since we assume that all the clients are disconnected at the beginning of each time interval, this client is recorded as having visited only one distinct AP in the second 12 hour interval: AP 3. Therefore, on average, this client visits 1.5 distinct APs in a 12 hour period this day.

The results of this procedure for each day, for all of the clients are shown in Figure 20. We show that, on average, 113.24 clients visit five or more APs in one day. In contrast, only 4.9 clients visit 15 or more APs in a day. In general, fewer clients have longer sessions. Figure 21 details these results.

Figure 20: Number of distinct clients who visit a selected number of different access points during time intervals of different length

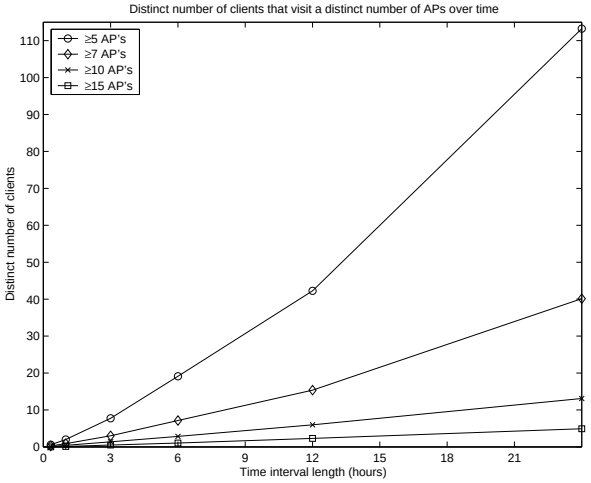


Figure 21: Average total and daily percent of distinct clients that visit a specified number of APs in a day

APs	≥ 2	≥ 3	≥ 5	≥ 15
$\mu_{clients}$	637.8	355.7	113.2	4.9
$\mu_{daily\%}$	45.0	23.8	7.2	0.3
$\sigma_{clients}$	336.0	226.6	85.9	4.1
$\sigma_{daily\%}$	10.4	9.6	4.3	0.21
$\frac{\lvert C.I._{95\%}\rvert}{2}$	143.7	96.9	36.7	1.75
$\frac{\lvert C.I._{95\%}\rvert}{2}_{daily\%}$	4.4	4.12	1.8	0.09

6.3 Session durations in access points

In this section we measure the number of different clients that remain connected to the access point for a given length of time. For this measurement we consider all of a client's sessions during a day. If a client has at least one session of at-least a given duration with an AP, we add that client to that AP's list of clients for that session duration. At the end of each day we count the number of unique clients in each of the AP's session duration lists.

Not surprisingly, the average number of clients that remain connected to an AP for a given time is lower for longer lasting sessions than for shorter lasting sessions. We find that on average, an AP has 8.55 distinct clients during the course of the day who have at least

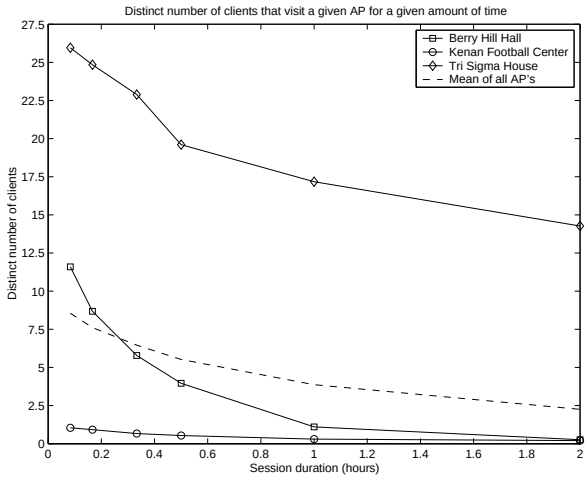


Figure 22: Clients with at least one AP session of the specified duration.

one session of five minutes or more. We also find that on average, an AP only has 2.26 distinct clients per day who have at least one session that lasts 2 hours or more. The standard deviations and 95% confidence intervals are $\sigma = 14.16$, $C.I._{95\%} = 8.55 \pm 1.92$ and $\sigma = 4.00$, $C.I._{95\%} = 2.26 \pm 0.53$, respectively. We find that certain buildings are representative of some of the major trends we see in our trace. Those buildings are: Tri Sigma House (a student residential area) represents an area where many clients are likely to remain for a longer period of time than they would in a class room. The Kenan Football Center represents areas with a small number of clients, and Berry Hill Hall (Biomedical Engineering building) represents an area in between. Their results are shown in Figure 22 along with the mean of all the APs in our trace.

7 Conclusions and future work

We examined the web requests and movement patterns of clients on the campus of a major university. We find that each client frequently requests objects which it has requested within the past hour, and occasionally requests objects which had been requested by other nearby users within the past hour. After this initial period, there is relatively little direct correlation among nearby users' requests for the same web objects. This suggests that the benefits of pre-fetching or caching for other clients are real, albeit short lived.

There is a significant correlation among the categories of information requested by nearby mobile users, with 22% of categorized requests being for categories for which another nearby user has requested information within the past hour. This suggests that systems which offer substantially similar data in the absence of access to the requested data may provide significant utility. However, since less than 8% of

Overall object, web server, and category popularity among our population of mobile users is modeled well as a Zipf function. This suggests that the findings of results on a single web server [3] extend in principle to the general population of web sites.

Finally, our results show that 45% of the clients on a given day visit more than one AP, and only 24% visit two APs. Also, the average AP had 8.55 clients per day with at least one 5 minute session, but only 3.9 clients who had at least one 1 hour session. The average AP in an academic building has more clients initiate a session in the middle of the day than late at night, whereas an average AP in a residential building has the most visits around midnight.

This work points to some clear open problems. We are working currently to develop a model predicting visitor patterns; existing techniques such as those described by Bhattacharya and Das [5] appear to require significant extension to perform well for mobile 802.11 users on a campus. To this end, we are investigating the use of a weighted difference between the observed steady-state probability vector and a Markov-chain based predictor. We are also considering computing a diffeomorphism, or a space warp, between these two vectors and measure the energy of that warp. We are also interested in determining if the clients session durations follow any known distribution.

The peer-to-peer caching systems which motivated this study initially, such as 7DS, require that objects be cacheable. Stale objects should not be distributed, but many popular objects on the web are not cacheable by the HTTP standard [8]. It appears that content providers use cacheability to force reloads of their pages for reasons *other* than document freshness; e.g., they wish to count readers, or to distribute new advertisements. This use of the cacheability mechanisms works well enough in fully-connected environments, but is a limiting factor for weakly-connected systems as we describe here. We intend to address this issue of cacheability; ideally, an object should be cached only for its *true* useful lifetime, while content providers receive the feedback they need. Our intuition about conventional communication media suggests that the useful lifetime of an object is arbitrarily long, though a measure of its freshness is always required. (This is the reason that every page of a newspaper includes the date of publication.)

We have noted that the web is not the ideal testbed for the measurement of location-dependent services. We are currently implementing several systems, including a collaborative note-sharing system for use in presentations and also a location-sensitive map-editing system.

We will perform measurements of the spatial locality effects with these applications deployed.

8 Thanks

Thanks go to Jay Aikat for care and feeding of the trace-collection system, and to Jim Gogan and Todd Lane for setup of the AP log generation system.

References

[1] Cisco Aironet 350 access point specifications. `http://www.cisco.com/en/US/products/hw/wireless/`.

[2] Open directory project. `http://www.dmoz.org/`.

[3] Adya, A., Bahl, P., and Qiu, L. Analyzing the browse patterns of mobile clients. In *Proceedings of the First ACM SIGCOMM Workshop on Internet Measurement* (2001), ACM Press, pp. 189–194.

[4] Balachandran, A., Voelker, G. M., Bahl, P., and Rangan, P. V. Characterizing user behavior and network performance in a public wireless LAN. *ACM SIGMETRICS Performance Evaluation Review 30*, 1 (2002), 195–205.

[5] Bhattacharya, A., and Das, S. K. LeZi-update: An information-theoretic approach to track mobile users in PCS networks. In *Mobile Computing and Networking* (1999), pp. 1–12.

[6] Breslau, L., Cao, P., Fan, L., Phillips, G., and Shenker, S. Web caching and zipf-like distributions: Evidence and implications. In *INFOCOM (1)* (1999), pp. 126–134.

[7] Castro, P., Greenstein, B., Muntz, R. R., Kermani, P., Bisdikian, C., and Papadopouli, M. Locating application data across service discovery domains. In *Mobile Computing and Networking* (2001), pp. 28–42.

[8] Duska, B. M., Marwood, D., and Freeley, M. J. The measured access characteristics of World-Wide-Web client proxy caches. In *Proceedings of the 1997 Usenix Symposium on Internet Technologies and Systems (USITS-97)* (Monterey, CA, 1997).

[9] Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., and Berners-Lee, T. Hypertext transfer protocol – HTTP/1.1. Internet Engineering Task Force Request For Comments 2616, 1999.

[10] Goodman, D., Borras, J., Mandayam, N., and Yates, R. Infostations: A new system model for data and messaging services. In *Proc. of IEEE VTC* (Phoenix, Arizona, May 1997), pp. 969–973.

[11] IEEE/ANSI. *Standard 802.11*, 1999 ed.

[12] Kotz, D., and Essien, K. Analysis of a campus-wide wireless network. Tech. Rep. TR2002-432, Dept. of Computer Science, Dartmouth College, September 2002.

[13] National Institute of Standards and Technology. *Federal Information Processing Standards Publication 180-1: Secure Hash Standard*.

[14] Papadopouli, M., and Schulzrinne, H. Effects of power conservation, wireless coverage and cooperation on data dissemination among mobile devices. In *ACM Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc)* (Long Beach, California, Oct. 2001).