Inferring Relative Popularity of Internet Applications by Actively Querying DNS Caches

Craig E. Wills, Mikhail Mikhailov and Hao Shang

Computer Science Department

Worcester Polytechnic Institute

{**cew,mikhail,hao**}**@cs.wpi.edu**

ACM SIGCOMM Internet Measurement Conference

Miami, Florida
October, 2003

Introduction

Problem: How to assess the relative popularity of Internet applications such as the Web or network games.

Possible approaches:

- Popularity lists—The list may be biased by paid placements. It only includes the most popular. What about usage of less popular?
- Logs of activity—good, but for one population of users. They may not be available or difficult to obtain.

Our Approach

Track the use of Domain Name Server (DNS) lookups for the servers used by an Internet application—Web servers, game servers, data servers.

Do so by "poking" at the contents of Local Domain Name Servers (LDNSs) caches, which store information about what DNS lookups have recently been performed by users of the LDNS.

A poke checks whether or not a server name record is currently cached at the LDNS.

Available Information

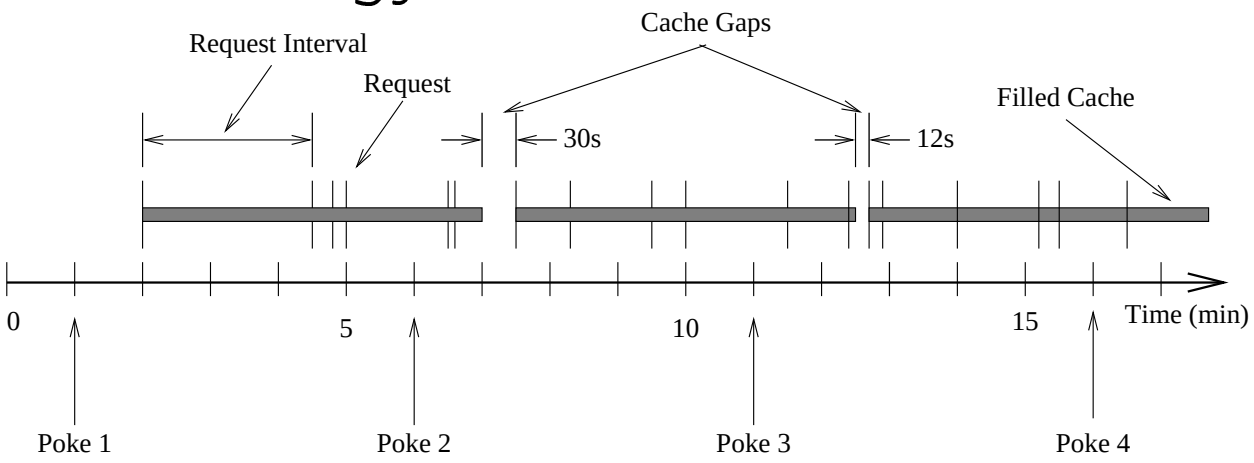
Can determine the relative popularity of a server name by measuring the

1. cache hit ratio of the name over time
2. cache gap interval for more popular servers to infer request arrival intervals.

Attractive because it can be applied to any Internet application that uses distinguished server names and performs DNS lookups on these names as part of use.

Cannot be used to measure precise usage information.

Methodology



- Periodically poke at the cache with non-recursive DNS queries, which do not pollute the cache.
- Use a period corresponding to the authoritative time-to-live (ATTL) for the server name (e.g. 5 min).
- Record the existence and the TTL of cached entries.

Analysis

- Compute the cache hit ratio over time.
- Can use the TTL value of successive pokes and the ATTL to determine the cache gap.
- Intuitively the smaller the cache gap, the more frequently the server is requested.
- If request interarrival times are exponentially distributed then the measured cache gaps will have the same distribution.

Comparison of Technique with Known DNS Requests

Obtained a log of DNS queries to the primary WPI DNS server for 28 continuous hours of mid-week activity in April 2003.

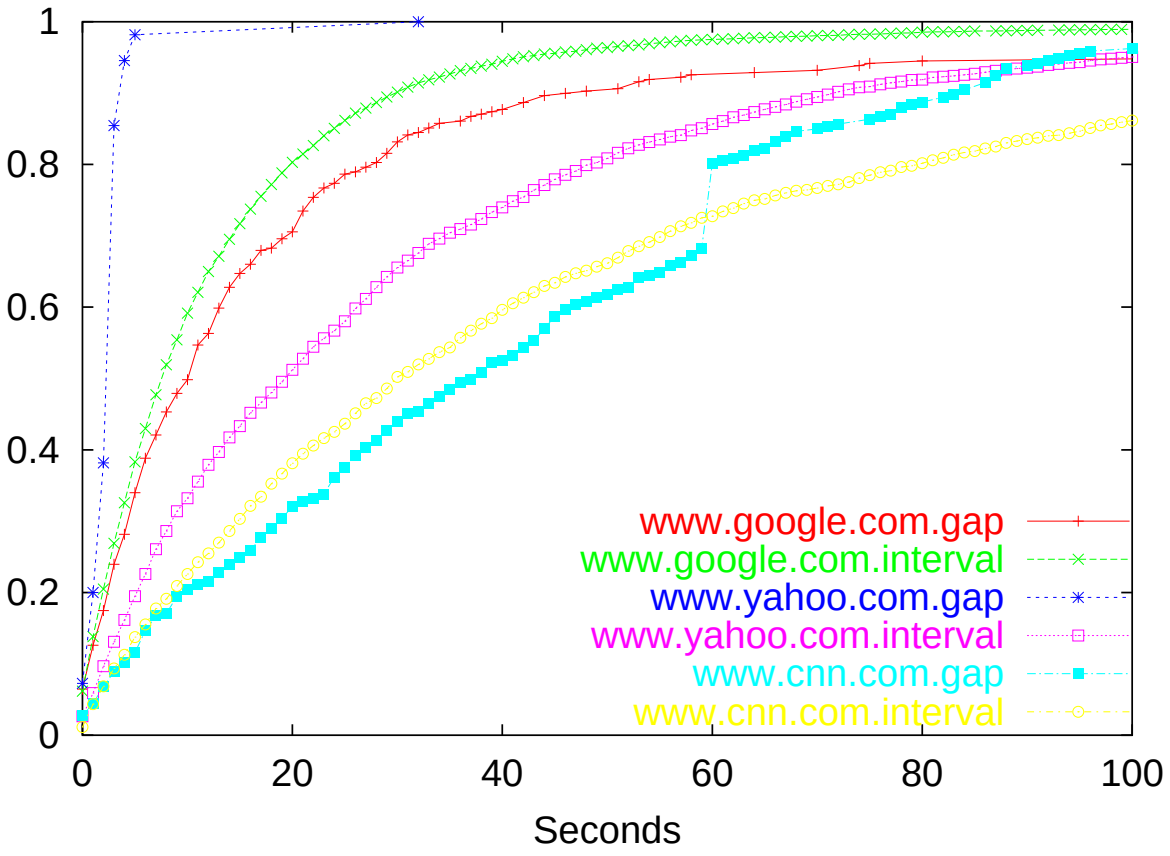
Focus on requests by WPI clients for non-WPI servers.

Applied our technique to this log to compare it with known request distribution.

Comparison Results

Use a Cumulative Distribution Function (CDF) to compare the measured Cache Gap with the known Request Interval.

	Cache Gap (sec)			Request Interval (sec)		
Server/ATTL	Med	Mean	StDev	Med	Mean	StDev
www.google.com/5m	11.0	25.5	56.0	8.0	14.3	25.7
www.yahoo.com/30m	3.0	3.0	4.1	20.0	32.0	42.3
www.cnn.com/5m	38.0	42.5	35.6	30.0	52.2	64.7



Good correspondence for most servers, but periodic application requests cause problems for the technique.

Potential Issues

- Cached records are flushed from a LDNS cache before they expire—a problem, but does not occur often and can be detected in analysis. Can also insert additional pokes.
- LDNS caches a non-authoritative record with a TTL less than ATTL. Similar problems and solutions as premature flushing of records.
- Potential denial-of-service (DOS) attack if queries are too frequent. At most we generated three requests per minute where WPI DNS server handles 5000 requests per minute.
- Privacy concerns. Could potentially correlate this technique with other data about users on a system.

Identification of Local DNS Servers

Generally found authoritative Domain Name Servers that also serve as LDNSs. Four categories with five LDNSs tested in each.

1. Commercial sites (com)
2. Educational sites (edu)
3. ISPs serving commercial sites (ispcom)
4. ISPs serving home customers (isphome)

Application Domains

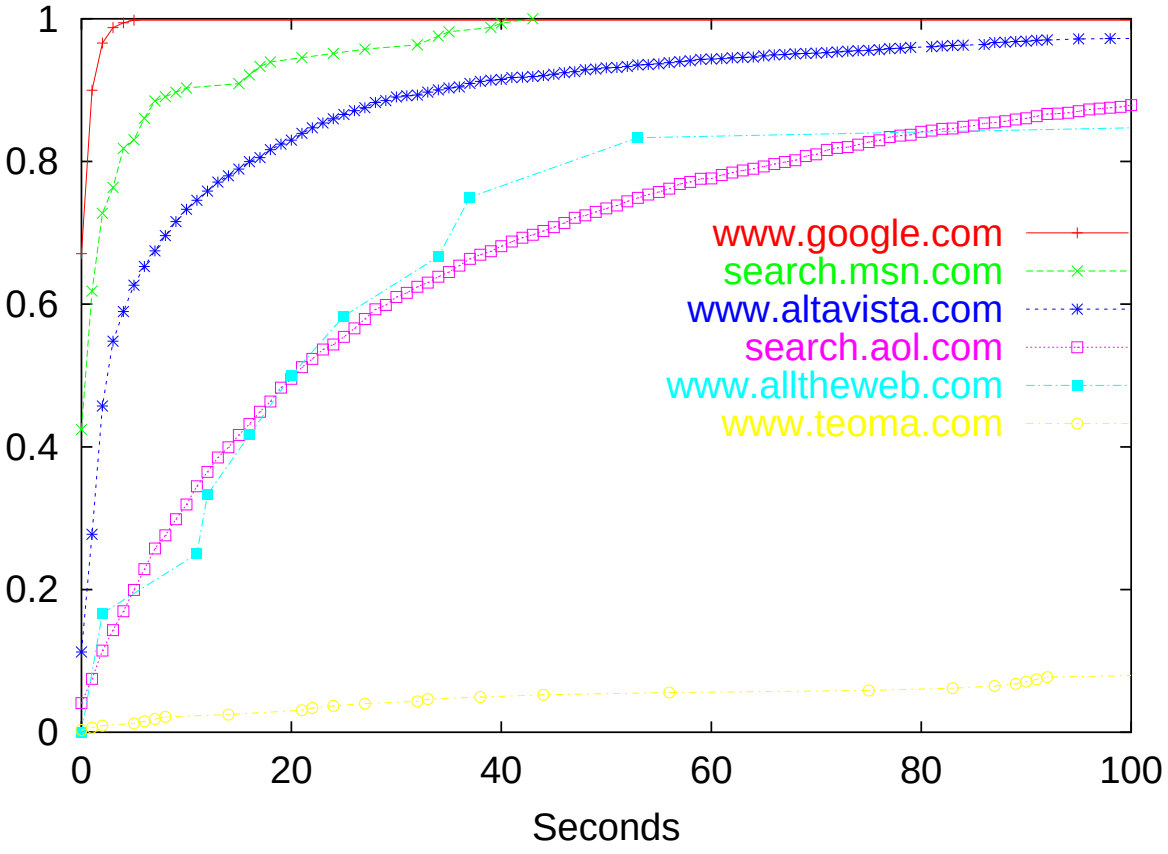
Can apply to any application that uses distinguished server names requiring a DNS lookup. All tests for approximately a one-week period.

Applications used in this study:

- Web search servers
- Web site traversal—`www.cnn.com` site
- Streaming content
- Network games—WarCraft/StarCraft, GameSpy
- Grid computing—distributed.net, seti@home

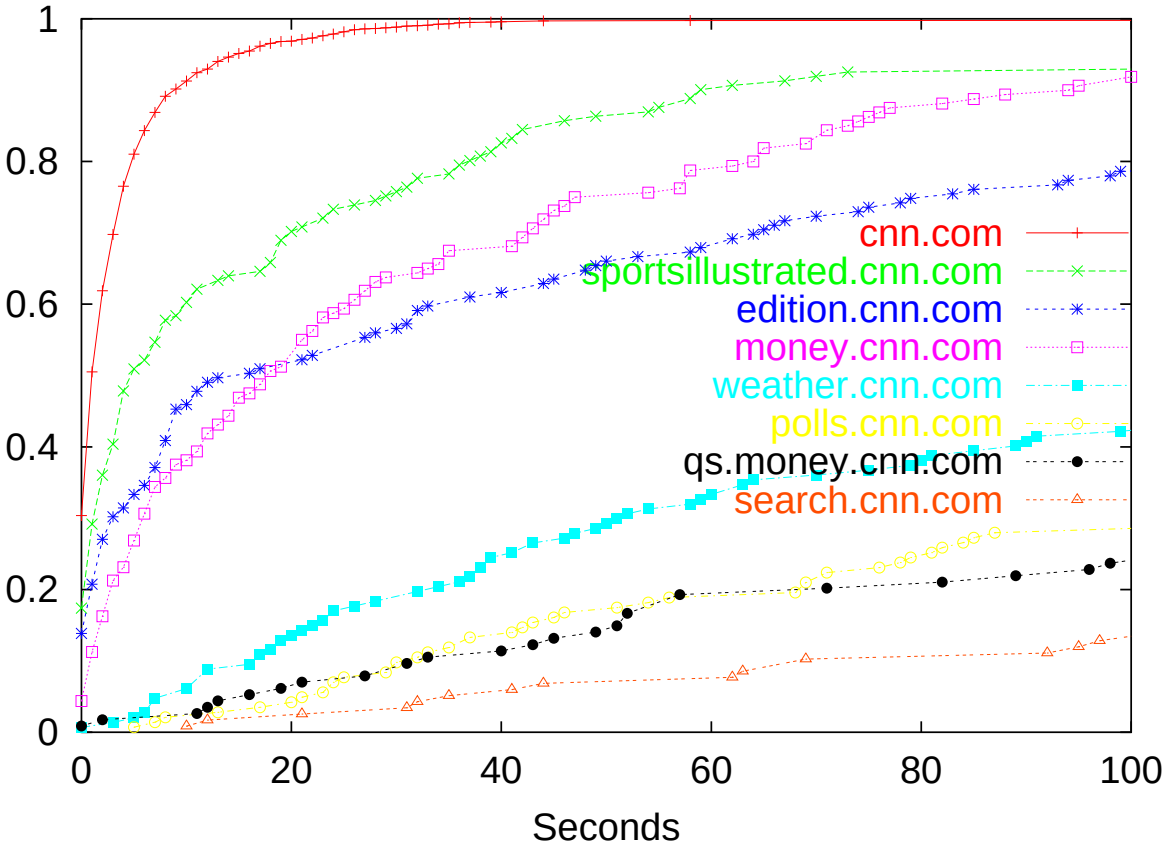
Web Search Servers for ispcom1

CDF of measured Cache Gap (sec):



CNN Web Site Traversal for ispcom1

CDF of measured Cache Gap (sec):



Network Games

Cache Hit % and median measured Cache Gap (sec):

Server/ATTL	Hit %	12h Hit %	Median
LDNS: edu5			
useast.battle.net/12h	92.9	92.9	921.5
uswest.battle.net/1h	14.5	76.9	14690.0
master.gamespy.com/1h	33.7	100.0	2316.0
LDNS: ispcom1			
useast.battle.net/12h	100.0	100.0	11.0
uswest.battle.net/1h	91.6	100.0	102.0
master.gamespy.com/1h	98.2	100.0	12.0
LDNS: isphome1			
useast.battle.net/12h	100.0	100.0	150.0
uswest.battle.net/1h	86.7	100.0	168.0
master.gamespy.com/1h	87.3	100.0	150.5

Summary

Can infer relative popularity of Internet applications by observing the presense of DNS records for distinguished server names.

Methodology can be employed for any server and at any LDNS for which access is available.

Cache hit percentage is a rough measure of popularity with the measured cache gap as an approximate measure of the request interval for more popular servers.

Future Work

- Other application domains—Instant messaging, CDNs
- Examine methodology for identifying LDNSs—possibly use LDNSs identified from DNS logs.
- Further validation of the methodology with known LDNS logs.
- Study inference of popularity from other types of caches such as Web caches.
- Study privacy exposure in using this approach.