
Cellular Data Network Infrastructure Characterization & Implication on Mobile Content Placement

Qiang Xu^{*}, Junxian Huang^{*}, Zhaoguang Wang^{*}
Feng Qian^{*}, Alexandre Gerber⁺⁺, Z. Morley Mao^{*}

^{*}University of Michigan at Ann Arbor

⁺⁺AT&T Labs Research



Applications Depending on IP Address

- ▶ IP-based identification is popular
 - ▶ Server selection
 - ▶ Content customization
 - ▶ Fraud detection

- ▶ Why? -- IP address has strong correlation with individual user behavior



This video is not available in your country.



Access Restricted (Bad IP)

You are trying to access Facebook from ;
with abusive behavior. You may request



Cellular IP Address is Dynamic

- ▶ Cellular devices are hard to geo-locate based on IP addresses
- ▶ One Michigan's cellular device's IP is located to places far away

IP2Location™ Live Product Demo

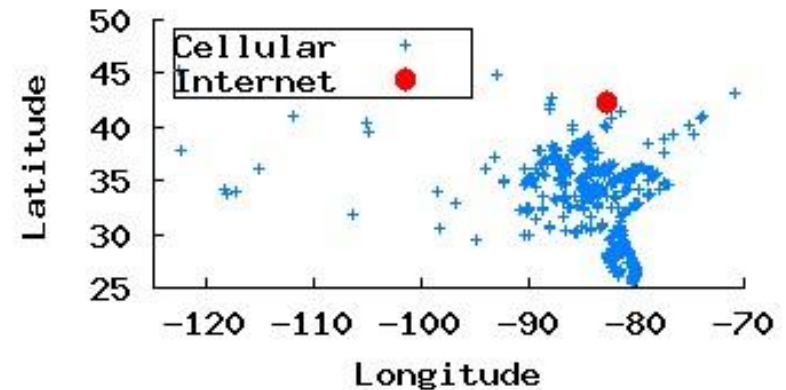
IP Address	Country	Region	City
	UNITED STATES	PENNSYLVANIA	DOYLESTOWN

MaxMind GeoIP City/ISP/Organization Edition Results

Hostname	Country Code	Country Name	Region	Region Name	City
166.137.136.51	US	United States	NY	New York	New York

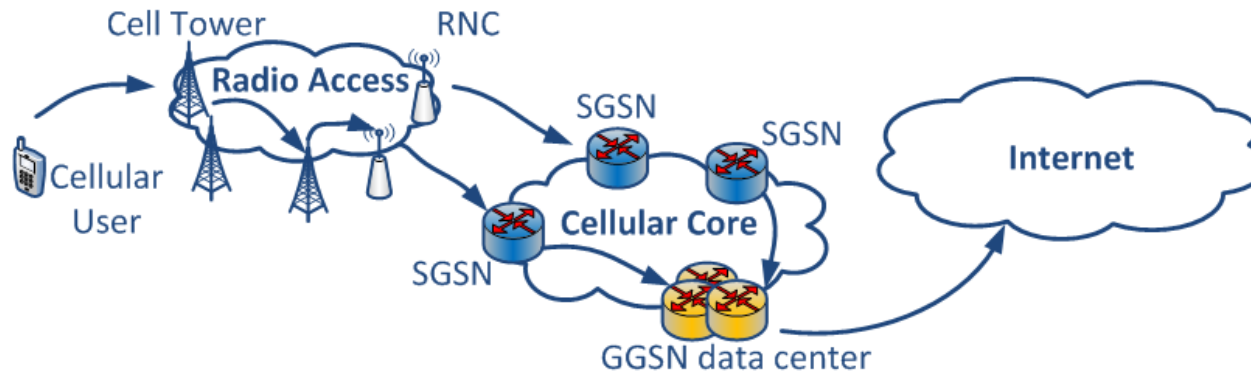
- ▶ /24 cellular IP addresses are shared across disjoint regions

Coverage of IP address



Problem Statement

- ▶ Discover the cellular infrastructure to explain the diverse geographic distribution of cellular IP addresses and investigate the implications accordingly



- * The first several IP hops are in GGSN data center
- * Cellular IP addresses are allocated by GGSN data center
- * GGSN data centers could be far away due to wireless hops

- ▶ The number of GGSN data centers
- ▶ The placement of GGSN data centers
- ▶ The prefixes of individual GGSN data centers

Challenges

- ▶ Cellular networks have limited visibility
 - ▶ The first IP hop (i.e., GGSN) is far away -- lower aggregation levels of base station/RNC/SGSN are transparent in *TRACEROUT*
 - ▶ Outbound *TRACEROUTE* -- private IPs, no DNS information
 - ▶ Inbound *TRACEROUTE* -- silent to ICMP probing
- ▶ Cellular IP addresses are more dynamic [BALAKRISHNAN *et al.*, IMC 2009]
 - ▶ One cellular IP address can appear at distant locations
 - ▶ Cellular devices change IP address rapidly

Solutions

- ▶ Collect data in a new way to get geographic coverage of cellular IP prefixes
 - ▶ Build Long-term and nation-wide data set to cover major carriers and the majority of cellular prefixes
 - ▶ Combine the data from both client side and server side

- ▶ Analyze geographic coverage of cellular IP addresses to infer the placement of GGSN data centers
 - ▶ Discover the similarity across prefixes in geographic coverage
 - ▶ Cluster prefixes according to their geographic coverage

Previous Studies

- ▶ Cellular IP dynamics
 - ▶ Measured cellular IP dynamics at two locations [Balakrishnan *et al.*, IMC 2009]
- ▶ Network infrastructure
 - ▶ Measured ISP topologies using active probing via TRACEROUTE [Spring *et al.*, SIGCOMM 2002]
- ▶ Infrastructure's impact on applications
 - ▶ Estimated geo-location of Internet hosts using network latency [Padmanabhan *et al.*, SIGMETRICS 2002]
 - ▶ On the Effectiveness of DNS-based Server Selection [Shaikh *et al.*, INFOCOM 2001]

Outline

- ▶ Motivation
- ▶ Problem statement
- ▶ Previous Studies
- ▶ **Data Sets**
- ▶ Clustering Prefixes
- ▶ Validating the Clustering Results
- ▶ Implication on mobile content placement

Data Sets

▶ DataSource1 (server logs): a location search server

- ▶ millions of records
- ▶ IP address, GPS, and timestamp

```

...
timestamp    lat.    long.    address
1251781217   36.75  -119.75  166.205.130.244
1251782220   33.68  -117.17  208.54.4.78
...
    
```

▶ DataSource2 (mobile app logs): an application deployed on iPhone OS, Android OS, and Windows Mobile OS

- ▶ 140k records
- ▶ IP address and carrier

```

device:
  <ID:C7F6D4E78020B14FE46897E9908F83B>
  <Carrier: AT&T>
address:
  <GlobalIP: 166.205.130.51>
...
    
```

▶ RouteViews: BGP update announcements

- ▶ BGP prefixes and AS number

```

... | 95.140.80.254 | 31500 | 166.205.128.0/17 | 31500 3267 3356 7018 20057 | ...
... | 95.140.80.254 | 31500 | 208.54.4.0/24 | 31500 3267 3356 21928 | ...
    
```



Map Prefixes to Carriers & Geographic Coverage

- ▶ Correlate these data sets to resolve each one's limitations to get more visibility

DataSource1

address	lat.	long.
166.205.130.244	36.75	-119.75
208.54.4.11	33.68	-117.17

RouteViews

prefix
166.205.128.0/17
208.54.4.0/24

DataSource2

address	carrier
166.205.130.51	AT&T
208.54.4.11	T-Mobile

prefix	lat.	long.
166.205.128.0/17	36.75	-119.75
208.54.4.0/24	33.68	-117.17

prefix	carrier
166.205.128.0/17	AT&T
208.54.4.0/24	T-Mobile

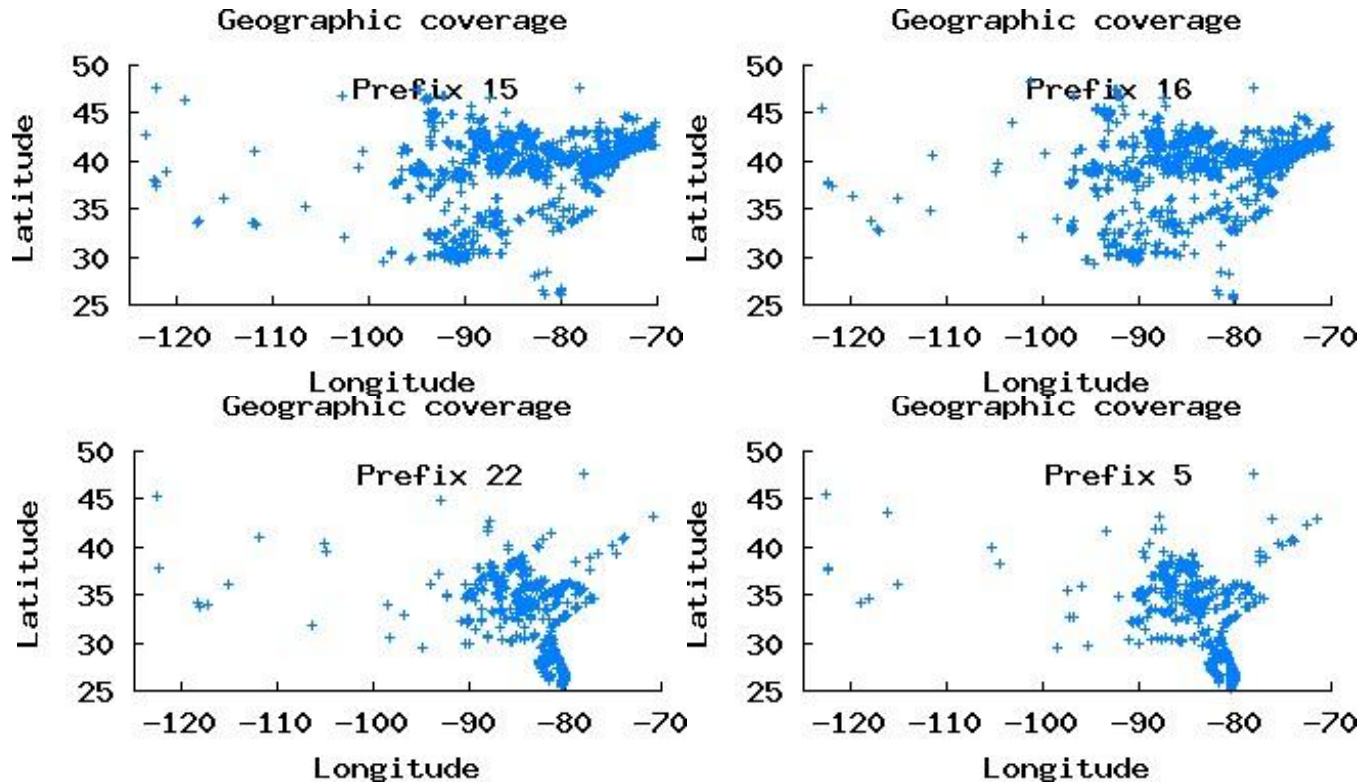
prefix	carrier	lat.	long.
166.205.128.0/17	AT&T	36.75	-119.75
208.54.4.0/24	T-Mobile	33.68	-117.17



Outline

- ▶ Motivation
- ▶ Problem statement
- ▶ Previous Studies
- ▶ Data Sets
- ▶ **Clustering Prefixes**
- ▶ Validating the Clustering Results
- ▶ Implication on mobile content placement

Motivation for Clustering -- Limited Types of Geographic Coverage Patterns



- ▶ Prefixes with the same geographic coverage should have the same allocation policy (under the same GGSN)

Cluster Cellular Prefixes

- ▶ 1. Pre-filter out those prefixes with very few records (todo)
- ▶ 2. Split the U.S. into N square grids (todo)
- ▶ 3. Assign a feature vector for each prefix to keep # records in each grid
- ▶ 4. Use bisect k-means to cluster prefixes by their feature vectors (todo)

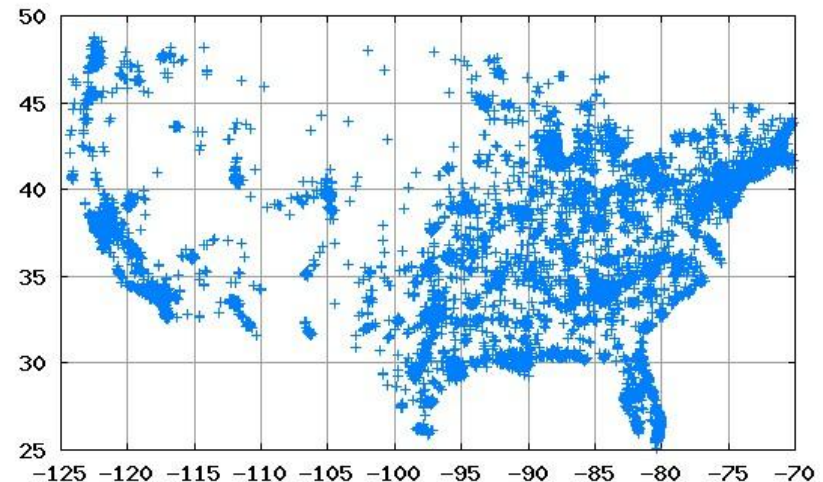
▶ How to avoid aggressive filtering?

- ▶ keep at least 99% records

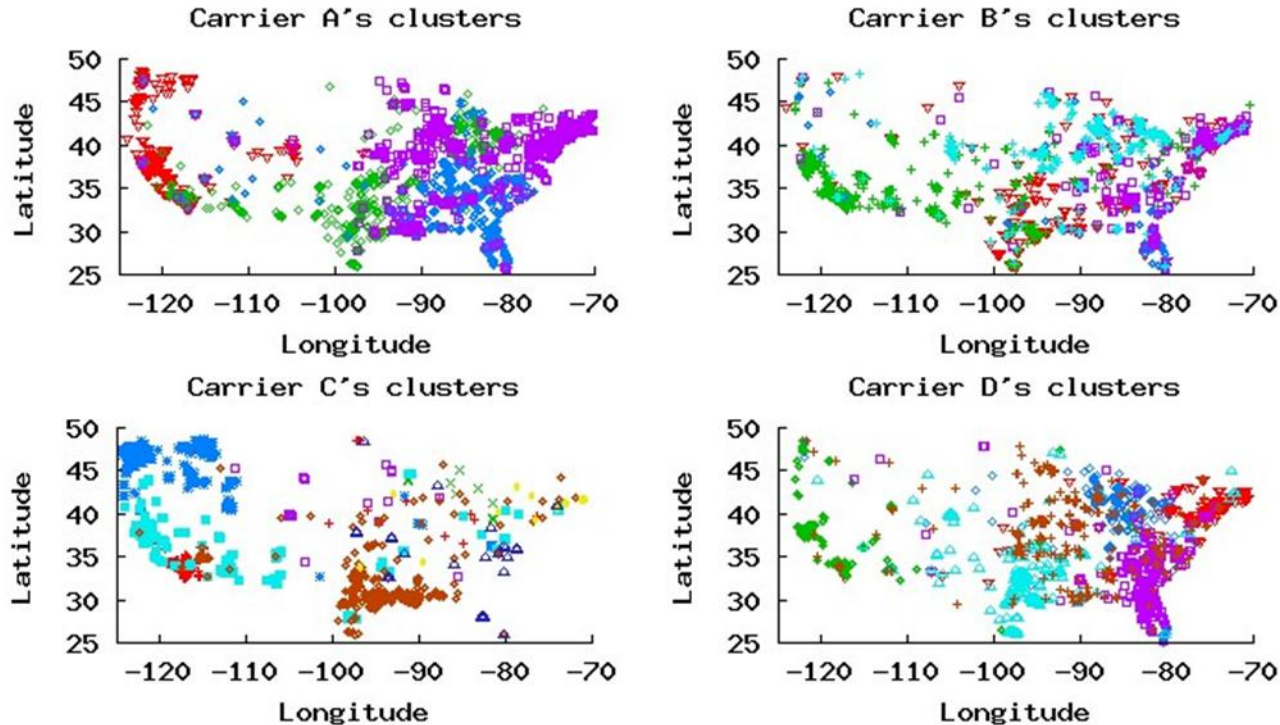
▶ How to choose N?

- ▶ # clusters is not affected by N while $N > 15$ && $N < 150$
 - ▶ The geographic coverage of each cluster is coarse-grained

▶ How to control the maximum tolerable SSE?



Clusters of the Major Carriers



All 4 carriers cover the U.S. with only a handful clusters (4-8)

- ▶ All clusters have a large geographic coverage
- ▶ Clusters have overlap areas
 - ▶ Users commute across the boundary of adjacent clusters
 - ▶ Load balancing

Outline

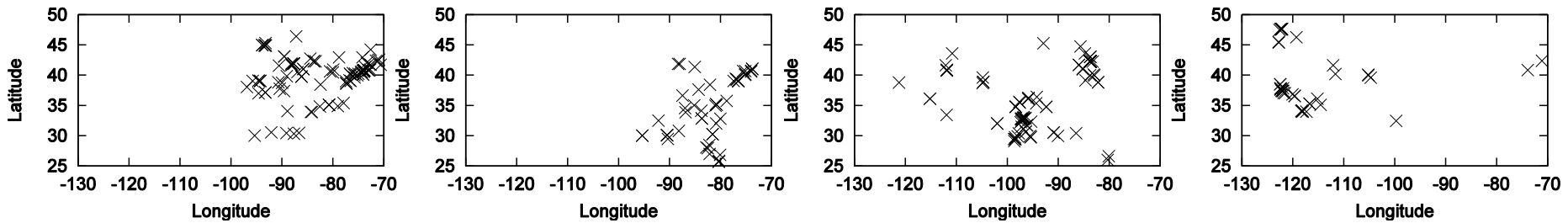
- ▶ Motivation
- ▶ Problem statement
- ▶ Previous Studies
- ▶ Data Sets
- ▶ Clustering Prefixes
- ▶ **Validating the Clustering Results**
- ▶ Implication on mobile content placement

Validate via local DNS Resolver (DataSource2)

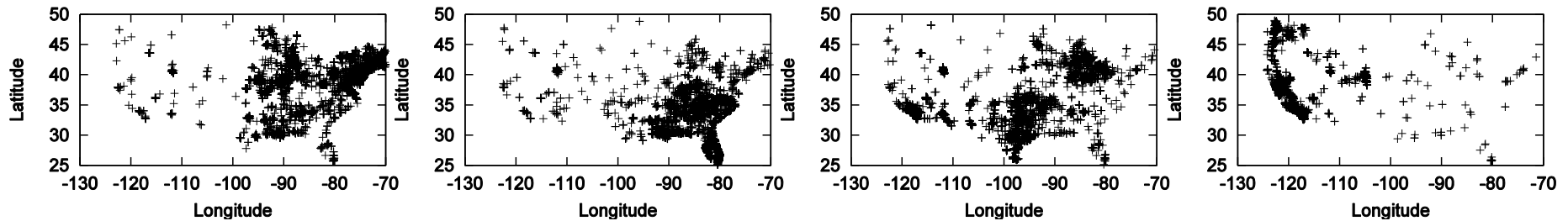
- ▶ Identify the local DNS resolvers
 - ▶ Server side: log the incoming DNS requests on the authoritative DNS resolver of **eecs.umich.edu** and record (id_timestamp, local DNS resolver)
- ▶ Profile the geographic coverage of local DNS resolvers
 - ▶ Device side: request **id_timestamp.eecs.umich.edu** and record the (id_timestamp, GPS)

Validate via Cellular DNS Resolver (Cont.)

► Clusters of Carrier A's local DNS resolvers



► Clusters of Carrier A's prefixes



Clustering Results

- ▶ Goal -- “...discover the cellular infrastructure to explain the diverse geographic distribution of cellular IP addresses...”
 - ▶ All 4 major carriers have only a handful (4-8) GGSN data centers
 - ▶ Individual GGSN data centers all have very large geographic coverage
- ▶ Goal -- “...investigate the Implications accordingly...”
 - ▶ Latency sensitive applications may be affected
 - ▶ CDN servers may not be able close enough to end users
 - ▶ Applications based on local DNS may not achieve higher resolution than GGSN data centers

Outline

- ▶ Motivation
- ▶ Problem statement
- ▶ Previous Studies
- ▶ Data Sets
- ▶ Clustering Prefixes
- ▶ Validating the Clustering Results
- ▶ **Implication on mobile content placement**

Routing Restriction:

How to Adapt Existing CDN service to Cellular?

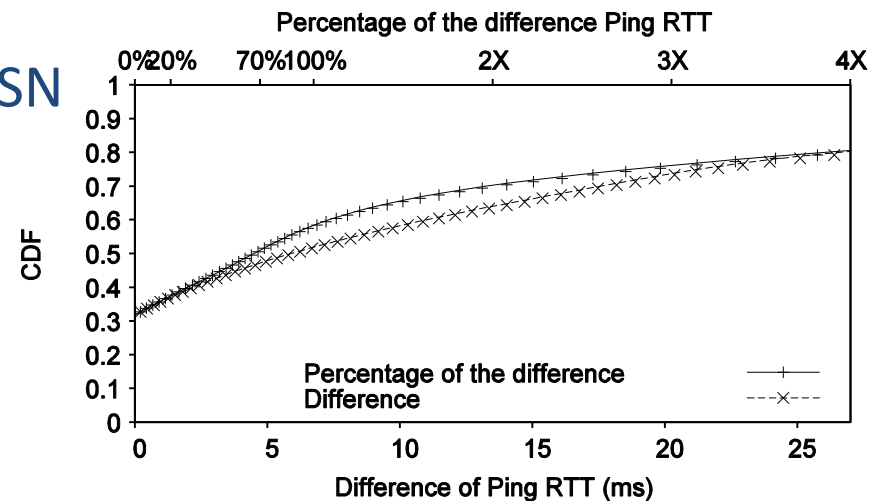
- ▶ Where to place content?
 - ▶ Along the wireless hops: require infrastructure support
 - ▶ Inside the cellular backhaul: require support from cellular providers
 - ▶ On the Internet: limited benefit, but how much is the benefit?
- ▶ Which content server to select?
 - ▶ Based on geo-location: finer-grained location may not available
 - ▶ Based on GGSN: location of GGSN

Server Selection (DataSource2)

- ▶ Approximately locate the server with the shortest latency
 - ▶ Based on IP address
 - ▶ Based on application level information, e.g., GPS, ZIP code, etc.

- ▶ Compare the latency to the Landmark server (1) **closest to device** with the latency to the Landmark server (2) **closest to the GGSN**
 - ▶ Estimate the location of GGSN based on *TRACEROUT*

- ▶ Select the content server based on GGSN!



Contributions

▶ Methodology

- ▶ Combine routing, client-side, server-side data to improve cellular geo-location inference
- ▶ Infer the placement of GGSN by clustering prefixes with similar geographic coverage
- ▶ Validate the results via *TRACEROUTE* and cellular DNS server.

▶ Observation

- ▶ All 4 major carriers cover the U.S. with only 4-8 clusters
- ▶ Cellular DNS resolvers are placed at the same level as GGSN data centers

▶ Implication

- ▶ Mobile content providers should place their content close to GGSNs
- ▶ Mobile content providers should select the content server closest to the GGSN

Thanks!

Questions?